Application of Penalized Splines Method in the Credit Market

Dissertation zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaften (Dr. rer. pol.) der Fakultät für Wirtschaftswissenschaften der Universität Bielefeld

vorgelegt von
M.A. Zhilin Yao

Bielefeld, 2008

Dekan:	Prof. Dr. Bernhard Eckwert
Gutachter:	Prof. Dr. Göran Kauermann
	PD. Dr. Pu Chen

Abstract

"Mortgage-related hit worse than expected has been a frequently cited phrase in recent months, and is usually followed by a list of victims consisting of banks and hedge funds. Although the current mortgage mess was caused by the subprime mortgage or bad credit mortgage, the broad impact of this subprime crisis promotes more concern in the loan lenders, the borrowers and mortgage-related products like mortgage backed securities. Since credit market is one of the successful application areas of statistics, we think a better understanding of MBS and its related risk and a more accurate credit risk assessment due to the application of more advanced statistics would contribute to the rebuilding of the credit market.

Nonparametric methods have proven to be useful in terms of capturing the flexible relationship between economic variables, while fewer assumptions about the economics constraint are made than in the traditional structural approach. This thesis follows the nonparametric trend and applies a new practical method, penalized splines, to investigate the issues on mortgage-backed securities (MBSs) and credit risk. The first application is to investigate the impacts of different interest rates on the prices of MBSs and show the hedging strategy based on the estimated smoothing functions. The second application concerns the stability of the impact of burnout effect on the prices of MBSs and its indication to the prepayment modelling. Finally, a credit risk model with varying coefficients is provided to explore the credit risk of small and medium-sized enterprises (SMEs) in China.

Acknowledgements

First and foremost I would like to thank my supervisor Prof. Dr. Göran Kauermann for his guidance throughout my doctoral studies. He introduced the nonparametric method in a summer mini-course and encouraged me to apply the penalized splines method to questions within the credit market. He has been extremely patient and a source of ideas throughout my studies. Without his guidance and support this PhD thesis would not have been written. A special thanks goes to PD. Dr. Pu Chen for many comments and introducing me to econometrics.

I also want to thank Prof. Dr. Volker Böhm and other Professors for the time and effort spent in the BiGSEM program. I am also grateful to Dr. Thorsten Pampel, Dr. Tomoo Kikuchi, Dr. Tatyana Krivobokova, Dr. Pavel Khomoski and other members of BiGSEM for the wide discussion, encouragement and friendship. Besides, I am also greatly indebted to Dr. Christoph Wöster and Dennis Kirchhoff for the kindly allowance to use the financial information system, Xtra 3000.

During my internship at the Commonwealth bank of Australia in Shanghai, I had the opportunity to get a taste of the financial world and found some interesting topics on mortgage-backed securities (MBSs) and credit risk modelling to explore. I would like to thank Anna, Annie, Daniel, Minhong, Steven and Yuan for the discussion and friendly time we spent together.

A very special thanks goes to my wife, Zhong, my daughter, Caipei and my parents for their love and support over the last years.

The financial support from DAAD, the Rotary-Clubs of Bielefeld and BiGSEM during my study is gratefully acknowledged.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Application of Nonparametric Methods in Finance

Financial markets are nowdays closely related to statistical theory. Campbell et al. (1997) describes this connection between financial models and statistical theory, *the random fluctuations that require the use of statistical theory to estimate and test financial models are intimately related to the uncertainty on which those models are based.* The break through idea of Black and Scholes (1973) and Merton (1973) promoted the usage of more sophisticated financial models in some financial applications such as pricing derivatives, modelling term structure of interest rate and credit risk modelling. As a result using different methods to calibrate these models have become popular, but most of the models to be calibrated are based on parametric functional relations between variables. Meanwhile, the nonparametric methods have been rapidly developed in the last two decades. Due to few assumptions about the underlying data structure, nonparametric methods are widely used in each area in the finance. Fan (2005) provides an overview of the application of nonparametric methods to financial econometrics. Here we use option pricing as an example to illustrate how nonparametric methods are used in finance area along with the development of nonparametric method itself. As in Aït-Sahalia and Lo (1998), the price of a European call option can be written as

$$C_i = g(S_i, K_i, T_i, r_i, \sigma_i) + \varepsilon_i \tag{1.1}$$

where C_i, S_i, K_i, T_i, r_i and σ_i are observations of call prices, stock prices, exercise prices, time to maturities, interest rates and volatilities respectively. Once the function form g is estimated, the state price density can be calculated by the following formula based on Banz and Miller (1978) and Breeden and Litzenberger (1978)

$$\hat{f}(S_T; t, T) = \frac{\partial^2 \hat{g}}{\partial K^2} \exp(r_{t,T}\,(T-t)) \tag{1.2}$$

Aït-Sahalia and Lo (1998) use the Nadaraya-Watson kernel estimator to estimate g in (1.1) and then calculated the second derivatives of $\hat{g}$ to substitute into (1.2). The Nadaraya-Watson estimator is also called local constant weighted estimator, proposed in Nadaraya (1964) and Watson (1964). Another widely used nonparametric method is the local polynomial model, which can be seen as the general model of the Nadaraya-Watson kernel estimator and the local linear estimator. Details can be found in Fan and Gijbels (1996). Aït-Sahalia and Duarteb (2003) make use of the local polynomial method to find the risk-neural measure using (1.2) and price options. Härdle and Yatchew (2002) apply the spline method together with a penalty to option pricing following (1.2). From this example, we see the development of the nonparametric approach. Other applications in finance reflecting the idea of fitting with a penalty go back to Rubinstein and Jackwerth (1996) and Lagnado and Osher (1997). Different from some of the above penalties, which have economic meaning, penalties can also be used to penalize the smoothness of the nonparametric unknown function in regression analysis, in which case this approach is called penalized splines. The application of penalized splines in the finance area can be observed in Jarrow et al. (2004), Kawasaki and Ando (2005), Krivobokova et al. (2006) and Wegener and Kauermann (2008). Penalized splines approach has the advantage of spline methods, the computational advantage of low-rank smoothers and the link to mixed models. This part will be discussed in Chapter 2.

1.2 Structure of the Thesis

Penalized splines have become a popular smoothing technique over the last couple of years. Originally introduced by O'Sullivan (1986), it was Eilers and Marx (1996) who demonstrated the simplicity and efficiency of the technique. Its link to a mixed model framework has been demonstrated in Ruppert et al. (2003). The link to mixed

model provides penalized splines with the ability to deal with correlation in error terms. This link makes it a more attractive tool among the nonparametric methods for financial data since a well-known feature of financial data is that they are usually serially correlated. Meanwhile, the process of globalization, the emergence of new economic powers such as in China and India and the technological improvement have had an impact on the evolution of the financial market. The banking sector is now actually facing a varying operation environment. This challenge also spurs the demand for new statistical methods, which can answer whether the change of basics exists and how to incorporate the potential changes into modelling.

The purpose of this thesis is to investigate the application of penalized splines smoothing to mortgage-backed securities and credit scoring in the banking industry. It begins with a theoretical introduction to penalized spline smoothing techniques. The main part of this thesis is concerned with three financial applications. Chapter 3 is devoted to the analysis of the impact of different interest rates on the prices of MBS. Although parametric models are popular, nonparametric techniques are also applied to problems related to MBS. Boudoukh et al. (1995), Boudoukh et al. (1997), LaCoure-Little et al. (1999), and Maxam and LaCour-Little (2001) demonstrate the application of kernel based nonparametric approach to pricing and prepayment modelling. To avoid the curse of dimensionality suffered by the kernel approach, Jegadrrsh and Ju (2000) model the prepayment rate using another nonparametric approach, the generalized additive model (GAM). We follow this nonparametric trend to explore the impact of different interest rates by using penalized splines smoothing, as recent powerful smoothing techniques. We first consider modelling the prices of a special MBS portfolio with similar maturities as nonparametric unknown functions of short term and long term interest rates together with random intercepts and correlated errors. Then we estimate these unknown functions and consider the use of the derivatives estimated by the nonparametric methods for hedging purpose.

Chapter 4 is concerned with the impact of burnout effect on the prices of MBS. Most reduced form prepayment models use constant coefficients for the explanatory variables in hazard models and model burnout independently. The coefficient sign of burnout is typically predicted to be negative according to economic intuition in advance and then verified by the estimation results later like in Schwartz and Torous (1993), Mattey and Wallace (2001) and Charlier and Bussel (2003). Kau

et al. (1992) and LaCour-Little and Green (2002) found that the coefficients are not constant. We illustrate the scheduled factor varying relationship between MBS prices and burnout effect. The result of this chapter indicates that the burnout effect has a different impact on the prices in different stages.

In the following chapter we introduce penalized splines smoothing to the credit risk assessment model. Chapter 5 deals with a credit risk assessment model with varying time effect for small and medium size enterprises (SMEs) in China. Among the consumer credit scoring methods the logistic model is quite popular due to its good balance between simplicity and accuracy. Altman and Sabato (2007), Behr and Guttler (2007) and Phillips and Vanderhoff (2004) show its application to predict loan default. We introduce a generalized smoothing model to credit scoring by the following procedures. We first assume a baseline risk related to time and then extend the commonly used constant coefficients assumption to a logistic model with time-varying coefficients. Finally, we estimate these unknown smoothing functions by using the penalized splines method. We also perform the out-of-sample validation and the out-of-time validation to illustrate the superiority of the model with time varying effect by comparing the results with other modelling strategies.

The common theme throughout all chapters of this thesis concerns the empirical application of the P-spline method. Most models used in the credit market are based on linear and constant coefficients assumptions. The main result of this thesis justifies the realization that these models are not flexible enough in order to exhaustively characterize the data. Even though this thesis is purely empirical, it is hoped that these results will broaden the insight into theoretical and empirical work.

Chapter 3 and Chapter 5 are in great part based on the joint work with Professor Kauermann. He provides ideas, discussion and suggestions for the following papers,

- Yao, Z. and Kauermann, G. (2008). Exploring the Credit Risk for Small and Medium-sized Enterprises in China. (submitted to *The Journal of Credit Risk*)

- Yao, Z. and Kauermann, G. (2008). Exploring the Impact of Different Interest Rates on the Prices of MBS using Penalized Splines. (submitted to *Real Estate Economics*)

- Yao, Z. (2008). Investigating the Burnout Effect of of MBS using Penalized Splines.

Chapter 2

Penalized Spline Smoothing and Its Extension

Nonparametric methods have been developed in the last two decades and Kauermann (2006) provides a summary of the main nonparametric models. In this chapter we focus our attention on one of these nonparametric models, penalized splines [2.1]. Its link to a mixed model framework has been demonstrated in Ruppert et al. (2003). The advantage of penalized splines is that the penalty prevents overfitting automatically and the estimation can be easily achieved due to its link with mixed models. We first introduce the penalized spline method and its extensions before the application part is discussed.

2.1 Penalized Splines Smoothing

Given scatterplot data (x_i, y_i), $1 \leq i \leq n$, the traditional way to summarize the relationship between x_i and y_i is to fit a linear model. In smoothing framework, we extend this linear restriction and assume that there is a smoothing functional relationship between x_i and y_i as follows,

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \cdots, n \tag{2.1}$$

[2.1] See O'Sullivan (1986) or Eilers and Marx (1996)

where $f(x_i) = E(y_i|x_i)$, is a unknown smoothing function, and ε_i are independent normally distributed residuals. To estimate the smoothing function $f(\cdot)$ we replace it by a linear combination of high dimension basis functions such that the nonlinearity can be captured by $\hat{f}(\cdot)$, the basis could be B-spline basis, truncated polynomial or radial basis e.t.c.[2.2] For the sake of simplicity, we illustrate the idea of penalized spline using linear basis like,

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k (x - \kappa_k)_+ \tag{2.2}$$

where β_0, β_1 and u_k are coefficients to estimate and $\kappa_1, \ldots, \kappa_k$ are fixed knots. The location of knots κ_k are chosen either as sample quantiles or equidistantly. Ruppert (2002) suggests $min\{n/4, 40\}$ as a choice for the number of knots. Instead of fitting (2.1) by least squares P-spline considers penalized least squares as follows,

$$Min \quad \|y - X\beta - Zu\| + \lambda u^T D u \tag{2.3}$$

where $\beta = (\beta_0, \beta_1)$, $X = (1, x)^T$, $Z = [(x-\kappa_1)_+, \ldots, (x-\kappa_k)_+]$, $u = (u_1, \ldots, u_k)^T$ $\lambda \geq 0$, is a smoothing parameter and D is a penalty matrix, which is an identity matrix in case of truncated polynomials basis.[2.3]

The solution of (2.3) and fitted values are,

$$\hat{\theta}_\lambda = (C^T C + \lambda D)^{-1} C^T y \tag{2.4}$$

$$\hat{y}_\lambda = C(C^T C + \lambda D)^{-1} C^T y \tag{2.5}$$

where $\theta = (\beta^T, u^T)^T$, $C = [X, Z]$, λ controls the tradeoff between the goodness of fit and the roughness. Figure 2.1 shows three penalized spline regression fits for λ values of 1, 0.1311 and 0.002. 25 knots are used. We estimate the true function $sin(3.14x_i/0.5) + 2$ using 100 observations generated from $y_i = sin(3.14x_i/0.5) + 2 + \varepsilon_i$ for 100 x_i in [0,1] and $\varepsilon_i \sim N(0, 0.16)$. The case λ = 0.1311 corresponds to a very satisfying fit. If we take λ to be larger, then the fit is similar to linear regression, as shown in Figure 2.1 (a). For λ = 0.002 we have decreased the penalty, so the fit is rougher. The large difference between these three fits illustrates the impact of λ

[2.2] The discussion of basis see Section 2.1.2

[2.3] See Ruppert et al. (2003) for the choice for penalty matrix of other basis.

on the fitting. Hence, the need for smoothing parameter selection arises if we want to decide on the amount of smoothing.

2.1.1 Data Driven Smoothing Parameter Selection

As we discussed earlier, the performance of the penalized smoothing depends on the choice of the smoothing parameter λ. To select the smoothing parameter automatically, we first require the specification of appropriate error criteria for measuring the error of the penalized smoothing at a single point as well as the error over the whole sample.

Let $\hat{f}_\lambda$ be an estimate of the function in (2.1). Define the mean squared error (MSE) at x_i by

$$MSE(\lambda) = E((y_i - \hat{f}_\lambda(x_i))^2)$$

It is clear to see that the MSE represents the bias-variance trade-off when it is decomposed into two components,

$$MSE(\lambda) = Var(\hat{f}_\lambda(x_i)) + (E(y_i - \hat{f}_\lambda(x_i)))^2$$

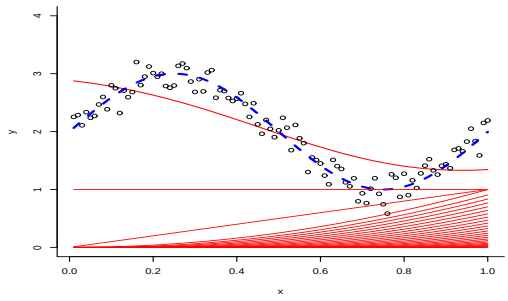
To choose the smoothing parameter in a global sense instead of only at the point x_i, we consider the average mean squared error,

$$AMSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E(y_i - \hat{f}_\lambda(x_i))^2$$

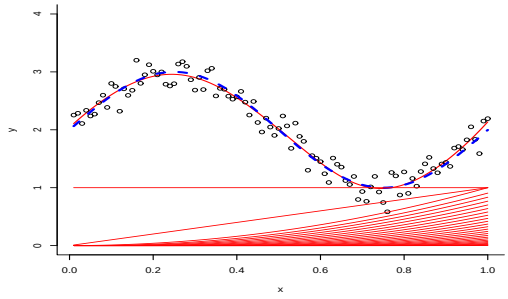
Another criterion that measures the performance of a model's prediction power is the average predictive squared error (APSE)

$$APSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E(y^* - \hat{f}_\lambda(x_i))^2 = AMSE(\lambda) + \sigma^2$$

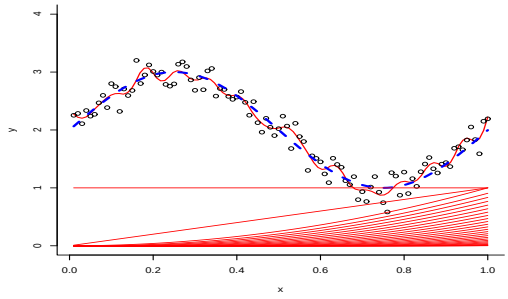
where $y_i^* = f(x_i) + \varepsilon_i^*$ is a new observation at x_i, ε_i^* are independent of ε_i and identically distributed with mean 0 and variance σ^2. APSE differs from AMSE by a constant σ^2. Let us denote $S_\lambda = X(X^TX + \lambda D)^{-1}X^T$ as a smoothing matrix, set $RSS(\lambda) = \|y - \hat{y}_\lambda\|^2$ and $df_{fit}(\lambda) = tr(S_\lambda)$. Then we have AMSE and APSE for linear smoother as follows



(a) λ=1



(b) λ=0.1311



(c) λ=0.002

Figure 2.1: The impact of smoothing parameter selection on the fitting based on 25 knots. The blue dashed line represents the true function while the upper solid red line is the penalized spline fit with the lower red basis

$$AMSE(\lambda) = \frac{tr(S_\lambda S_\lambda^T)}{n}\sigma^2 + RSS(\lambda)/n$$

$$APSE(\lambda) = [1 + \frac{tr(S_\lambda S_\lambda^T)}{n}]\sigma^2 + RSS(\lambda)/n$$

The bias-variance trade-off can be easily seen from the above equations. As shown in Figure 2.1, the bias increases as the amount of smoothing increase while the variance decreases, and vice versa. We can use residual sum of squares (RSS) as a measure of how well the spline fits observations. But only minimizing RSS over the smoothing parameter will result in an interpolating estimate. Therefore, we aim to find an optimal smoothing parameter that compromises between the goodness of fit and model complexity.

There are many model selection methods which can be used to select a smoothing parameter, such as Akaike's AIC (Akaike 1973), Mallow's C_p (Mallows 1973), cross-validation (CV) (Stone 1974), and generalized cross-validation (GCV) (Craven and Wahba 1979). We will briefly introduce these most commonly used methods.

Cross-Validation

Unlike RSS, which uses the same sample for model fitting and model evaluation, cross-validation uses the principle of "leave-one-out prediction. The idea is to leave the data points out one at a time and to select the smoothing parameter under which the removed data points can be best predicted by the remaining data. The cross-validation criterion to be minimized is as follows,

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_{-i}(x_i, \lambda)^2$$

where $\hat{f}_{-i}(x_i, \lambda)$ is the estimation based on leaving out data point (x_i, y_i). Craven and Wahba (1979) show that CV can also be represented as,

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{f}(x_i, \lambda))^2}{(1 - S_\lambda^{ii})^2} \tag{2.6}$$

where $\hat{f}(x_i, \lambda)$ is the fit using all data points, S_λ^{ii} is a diagonal element of S_λ. The CV in expectation approximates APSE if we assume $H_{ii} \approx \{SS^T\}_{ii}$, $1/(1-H_{ii})^2 \approx 1 + 2H_{ii}$ and $b_i = y_i - \hat{f}(x_i, \lambda)$, we have

$$E\{CV(\lambda)\} = APSE(\lambda) + \frac{1}{n}\sum_{i=1}^{n} H_{ii}(\lambda) b_i^2(\lambda)$$

Mallows's C_p

Mallows's C_p approximates APSE by adding $2tr(S_\lambda)\sigma^2/n$ to RSS/n. For unknown σ^2 we use an estimator $\hat{\sigma}^2$

$$C_p(\lambda) = \frac{2tr(S_\lambda)}{n}\hat{\sigma}^2 + RSS(\lambda)/n \tag{2.7}$$

where

$$\hat{\sigma}^2 = \frac{RSS(\lambda)}{n - tr(2S_\lambda - S_\lambda S_\lambda^T)}$$

Generalized Cross-Validation

Replacing S_λ^{ii} in CV by the average of all diagonal elements $tr(S_\lambda)/n$, Craven and Wahba (1979) also propose the following generalized cross-validation

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat{f}(x_i, \lambda))^2}{(1 - tr(S_\lambda)/n)^2} = \frac{RSS(\lambda)/n}{\{1 - tr(S_\lambda)/n\}^2}$$

Here we could also find the difference between RSS and GCV. The probability that GCV(λ) select $\lambda = 0$ (interpolation) is non-zero, otherwise $tr(S_\lambda) = n$ and the denominator is zero. GCV is a weighted version of CV with weights $(1 - S_\lambda^{ii})^2/(1 - tr(S(\lambda)/n)^2$. Moreover, if $tr(S_\lambda)$ is small, using the approximation $(1-x)^{-2} \approx 1 + 2x$,

$$GCV(\lambda) \approx \frac{1}{n}RSS + \frac{2tr(S_\lambda)}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i, \lambda))^2 \tag{2.8}$$

Comparing (2.8) to (2.7), we find that the GCV approximate C_p if RSS/n is regarded as an estimate of σ^2.

Akaike Information Criterion

Assuming the data y are generated from the distribution with true density $e_0(y)$ and $e_{\hat{\theta_\lambda}}(y)$ represents the density generating observations y for estimate $\hat{\theta_\lambda}$, then the Kullback-Leibler discrepancy of the estimator $\hat{\theta}_\lambda$ is defined as

$$K(e_{\hat{\theta_\lambda}}, e_0) = \int \{log[e_0(y)] - log[e_{\hat{\theta}_\lambda}(y)]\} e_0(y) dy$$

It can be shown that

$$E\{\hat{K}(e_{\hat{\theta_\lambda}}, e_0)\} = -l(\hat{\theta}) + p + \int log[e_0(y)] e_0(y) dy$$

where $-l(\hat{\theta}) = -log[e_{\hat{\theta_\lambda}}(y)]$, is the maximized log likelihood function. AIC is the double of the first two terms.

$$AIC = 2[-l(\hat{\theta}) + p] \tag{2.9}$$

In (2.9), the compromise takes place between the maximized log likelihood and p, the number of free parameters estimated within the model, which can be seen as a measure of complexity. If the model errors are normally and independently distributed, then we obtain

$$AIC(\lambda) = log\{RSS(\lambda)\} + 2df_{fit}(\lambda)/n$$

Following the idea of Hurvich and Tsai (1989), Hurvich et al. (1998) propose the following modified AIC for smoothing parameter selection corresponding to finite sample,

$$AIC_m(\lambda) = nlog\{RSS(\lambda)\} + \frac{n(2tr(S_\lambda) + 2)}{n - tr(S_\lambda) - 2}$$

2.1.2 Basis

To apply penalized splines, basis functions will be chosen so that a balance of numerical stability and practicality can be reached. Here we first briefly consider some popular choices for the basis functions in a one dimension case. The basis used in bivariate case will be discussed in Section 2.2.

Truncated polynomials

In addition to the truncated linear function used in (2.2), a truncated basis function of pth-degree is as follows,

$$1, x, \ldots, x^p, (x-\kappa_1)_+^p, \ldots, (x-\kappa_k)_+^p \tag{2.10}$$

and the corresponding spline is,

$$f(x) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{p} u_k (x-\kappa_k)_+^p \tag{2.11}$$

An example of truncated polynomial spline is shown in Figure 2.1. A benefit of choosing truncated polynomials is its ease in calculating its derivatives, which is used for some problems in finance such as the impact of interest rates discussed in Chapter 3. However, the truncated lines do not have optimal numerical properties [2.4].

B-spline

B-splines are numerically more stable than truncated polynomials. We define $m+1$ nondecreasing numbers, $k_0 \leq k_1 \leq k_2, \ldots, \leq k_m$, as knots. Then, the i-th B-splines function of order p is obtained by recurrence from first-order B-splines,

$$p = 1 \quad B_{i1}(x) = \begin{cases} 1 & \text{if } k_i \leq x < k_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

$$p > 1 \quad B_{ip}(x) = \frac{x-k_i}{k_{i+p-1}-k_i} B_{i,p-1}(x) + \frac{k_{i+p}-x}{k_{i+p}-k_{i+1}} B_{i+1,p-1}(x)$$

The whole process is shown in Figure 2.2. Figure 2.3 shows two examples. The properties of B-spline include, $\sum_{i=0}^{p} B_{ip}(x) = 1$, $B_{ip}(x) > 0$ if $k_i < x < k_{i+p}$, $B_{ip}(x) = 0$ $k_0 \leq x \leq k_i$, $k_{i+p} \leq x \leq k_{n+p}$ and $B_{ip}(x)$ is $(p-2)$ times continuously differentiable if knots are pairwise different from each other. Given a set of n+1 control points (called de Boor points), β_0, β_1, ..., β_n, and knots discussed above, a B-spline $f(x)$ of order p is defined as

[2.4] see Aerts et al. 2002

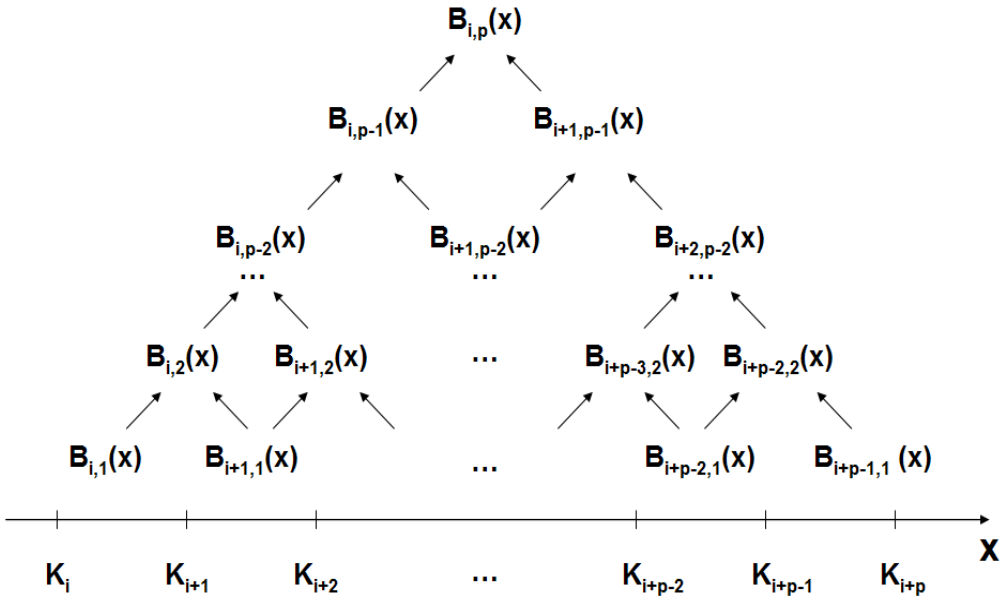


Figure 2.2: Generation of B-spline

$$f(x) = \sum_{i=0}^{n} B_{ip}(x)\beta_i$$

where $B_{ip}(x)$ is the B-spline function of degree $p-1$ based on the corresponding knots. The degree of the polynomial does not exceed $p-1$. The first $p-2$ derivatives are continuous.

Radial basis functions

A radial basis of pth degree is as follows,

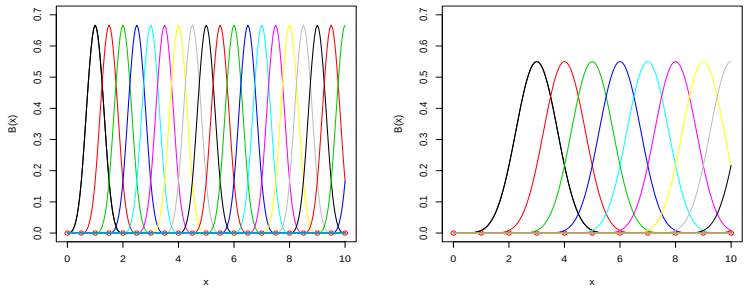


Figure 2.3: Examples of B-spline

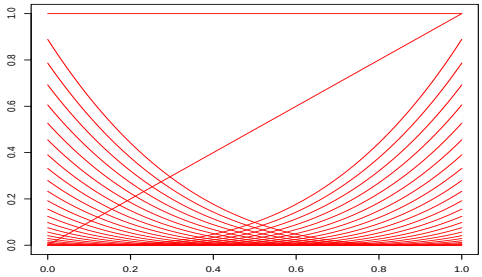


Figure 2.4: Radial basis functions of degree 3

$$1, x, \ldots, x^{\frac{p-1}{2}}, |x-\kappa_1|^p, \ldots, |x-\kappa_k|^p \qquad p = 1, 3, 5, \ldots \tag{2.12}$$

and the corresponding spline is,

$$f(x) = \beta_0 + \beta_1 x + \ldots + \beta_p x^{\frac{p-1}{2}} + \sum_{k=1}^{p} u_k |x-\kappa_k|^p \tag{2.13}$$

An example of radial spline is shown in Figure 2.4. Radial basis is numerically more stable and easy to implement in any number of dimensions. A two-dimension case can be found in Figure 2.6.

2.1.3 Number and Location of Knots

The first approach in choosing the number of knots is analogous to the model selection. The model selection criterion described in Section 2.1.1 can be used to choose the combination of knots automatically. Another approach, which is very simple, Ruppert (2002) investigates the impact of the number of knots on the performance of penalized splines using two algorithms and recommends a default that uses $K = min(n/4, 40)$. However, a researcher with some knowledge of the shape of the smoothing function may very well be able to select the number of knots without using an automatic algorithm.[2.5] This should be an important point as the penalized splines method is applied to empirical data in economics, because people always have more or less information about the relationship between economical variables. The location of the knots is usually chosen either as sample quantiles or equidistantly.

2.1.4 Link with Linear Mixed Model

Mixed models are models including random effect terms in addition to a random error term and fixed effect term. They are often used to cope with data with clustered structures, where the classical statistics assumption that observations are independent and identically distributed (iid) could fail. Linear mixed effect (LME) models may be viewed as a generalization of the variance component and regression models. Penalized likelihood is frequently used to cope with parameter multidimensionality. Penalized likelihood may be derived from a mixed model as an approximation to the marginal likelihood after applying the Laplace approximation. Moreover, the penalty coefficient, often derived from a heuristic procedure, is estimated by maximum likelihood as an ordinary parameter. Since the mixed model naturally leads to penalized likelihood, it can be applied to penalized smoothing. In particular, the difficult problem of selecting a smoothing parameter (penalty coefficient) selection can be solved by the mixed model technique by estimating this coefficient from the data.

Model (2.2) can be represented as a mixed model by treating the coefficients u as a random effect, thus the estimation of model (2.2) can be accomplished by using

[2.5] See Ruppert (2002), page 753

the well-developed algorithms and softwares for mixed models. A matrix form of model (2.2) like,

$$y = X\beta + Zu + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2 I), \quad u \sim N(0, \frac{\sigma_\varepsilon^2}{\lambda}\tilde{D}^{-1}),$$

where $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_u^2}$, is a variance ratio, $\tilde{D}^{-1}$ is a penalty matrix. More generally, we assume $\varepsilon \sim N(0, R)$, $u \sim N(0, G)$ and set $V = Cov(y) = ZGZ^T + R$. Hence y has a multivariate normal distribution $N(X\beta, V)$, then the loglikelihood of y is,

$$l(\delta, V) = -\frac{1}{2}\{nlog(2\pi) + log|V| + (y - X\beta)^T V^{-1}(y - X\beta)\} \tag{2.14}$$

Given that V maximizes log-likelihood function (2.14) we get the estimate of β,

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$$

For known β, G and R the random coefficients u can be predicted as following BLUP,[2.6]

$$\tilde{u} = GZ^T V^{-1}(y - X\beta)$$

For unknown G and R, the parameters in covariance matrices G and R can be estimated by the most widely used algorithms, maximum likelihood (ML) or restricted maximum likelihood (REML).[2.7] Then we obtain the estimated best linear estimate (EBLUE) and the estimated best linear predictor (EBLUP) as follows,

$$\hat{\beta} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} y$$

$$\tilde{u} = \hat{G} Z^T \hat{V}^{-1}(y - X\hat{\beta})$$

2.1.5 Correlated Errors

It is well known that the automatic smoothing parameter chooser such as GCV and AIC tend to undersmooth (positive correlation) or oversmooth (negative correlation) the data when correlated errors exist. Methods of smoothing with correlated

[2.6] See Robinson (1991).
[2.7] See Harville (1977) and Harville (1974).

errors have been proposed as in many applications such as in finance where the data are usually serially correlated. Opsomer et al. (2001) investigates several data driven smoothing parameter selection methods as well as regression models and shows that data driven smoothing parameter selection tends to undersmooth or oversmooth corresponding to cases of positive or negative correlation. Currie and Durbán (2002) found that a good estimate of the underlying smoothing function and the correlation coefficients can be achieved when REML is used. Durbán and Currie (2003) represent penalized spline smoothing as linear mixed model and use REML to estimate correlation parameters. Krivobokova and Kauermann (2007) show that REML outperforms AIC when the correlation structure is misspecified.

2.2 Bivariate Smoothing

So far we have only considered models with a single covariate, which only captures a function of one variable. In many cases, the interaction between variables is also of interest. So ideally, we want a two-dimensional function which captures both variables. Thus, a bivariate smoothing function is to be estimated. The general bivariate smoothing model as an extension of (2.1) is,

$$y_i = f(x_{1i}, x_{2i}) + \varepsilon_i \tag{2.15}$$

To estimate the unknown two dimensional function in (2.15), bivariate basis functions are required. Bivariate basis functions can be generated by two methods. One such method is taking products of two one-dimensional basis such as truncated power functions and B-spline, while the other is directly using radial basis functions which are defined as functions of the distances between the data and the knots. Suppose that we have two sets of one dimensional basis functions, $\mathcal{B}_1 = \{B_{x_{1j_1}} : j_1 = 1, 2, \ldots, K_{x_1}\}$ and $\mathcal{B}_2 = \{B_{x_{2j_2}} : j_2 = 1, 2, \ldots, K_{x_2}\}$, for x_{1i} and x_{2i} respectively. The tensor product of these two sets of basis functions is

$$\mathcal{B}_{12} = \mathcal{B}_1 \otimes \mathcal{B}_2$$

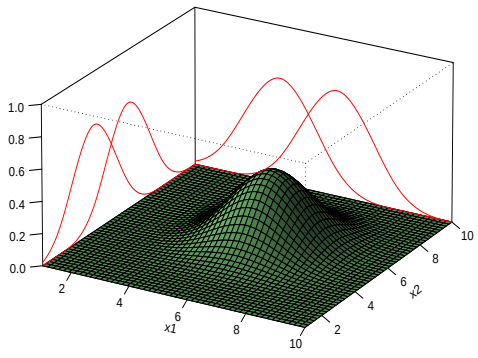


Figure 2.5: An example of tensor product basis function. Here we only show one surface.

The function $f(x_{1i}, x_{2i})$ in (2.13) is now modelled as follows,

$$f(x_{1i}, x_{2i}) = \sum_{j_1=1}^{K_{x_1}} \sum_{j_2=1}^{K_{x_2}} \beta_{j_1 j_2} B_{x_{1j_1}}(x_{1i}) B_{x_{2j_2}}(x_{2i}) \tag{2.16}$$

In case of taking products of basis functions such as in (2.16), the penalty part in (2.3) also extends to two penalties corresponding to two directions. An example of these basis functions is shown in Figure 2.5.

Radial basis functions of (x_{1i}, x_{2i}) are of the form

$$C(\|(x_{1i}, x_{2i}) - (k_{1j}, k_{2j})\|) \qquad j = 1, 2, \ldots, K$$

where $C(\cdot)$ is an univariate function as follows,

$$C(r) = \|r\|^{2m} log\|r\| \qquad m = 1, 2, \ldots \tag{2.17}$$

(2.17) are also called polyharmonic spline radial basis functions. $f(x_{1i}, x_{2i})$ in

(2.15) is now modelled by radial basis functions of form,

$$f(x_{1i}, x_{2i}) = 1 + \beta_1 x_{1i} + \beta_2 x_{2i} + \sum_{j_1=1}^{K} C(\|(x_{1i}, x_{2i}) - (k_{1j}, k_{2j})\|) \tag{2.18}$$

Figure 2.6 shows an example of radial basis functions.

2.3 Varying Coefficients Model

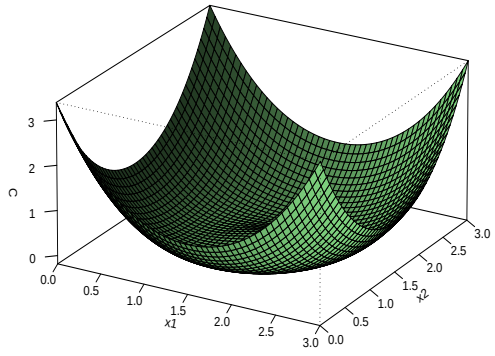
The varying coefficients model is introduced in Hastie and Tibshirani (1993), as an extension of the linear regression model. The varying coefficients model has different slopes while the linear assumption remains. Suppose that we have three variables(y_i, x_{1i}, x_{2i}), $i = 1, 2, \ldots, n$, then a varying coefficients model describes the structure of these data as

$$y_i = g(x_{1i}) + f(x_{1i})x_{2i} + \varepsilon_i, \tag{2.19}$$

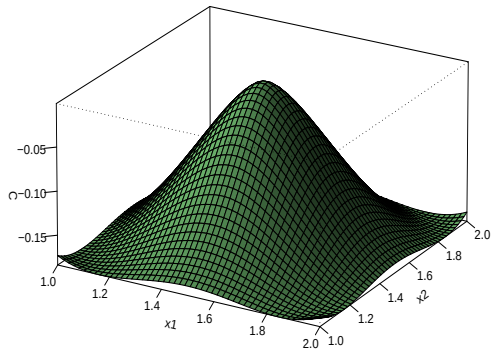
where f, a function representing the coefficients of x_{2i}, is assumed to change smoothly over x_{1i}. Substituting (2.2) into g and f in (2.19) respectively, we can rewrite a varying coefficients model as a penalized linear spline,

$$y_i = \beta_{10} + \beta_{11}x_{1i} + \sum_{k=1}^{K} u_{1k}(x_{1i} - \kappa_k)_+ + [\beta_{20} + \beta_{21}x_{1i} + \sum_{k=1}^{K} u_{2k}(x_{1i} - \kappa_k)_+]x_{2i} + \varepsilon_i, \tag{2.20}$$

where κ_k are knots chosen for x_{1i}. Fitting this model and linking it to a mixed model are standard as in Section 2.1. The extension of (2.19) includes replacing the univariate functions g and f with two dimensional surface functions in Rau et al. (2007). The varying coefficients model is useful in economics and financial application when we want to trace the effects of some variables, or to find out, when the effects come to play a role and how the effects vary over the observation period.



(a)



(b)

Figure 2.6: An example of 2-D radial basis function with (1.5, 1.5) as one of the knots. Plot (b) is the rescaled center part of plot(a).

2.4 Additive Model

Analogous to the extension of an univariate linear regression model to (2.1), the additive model is an extension of the multiple regression model. For one-dimensional response variables y_i and d-dimensional covariates $X_i = (x_{1i}, ..., x_{di})$, $i = 1, ..., n$, the additive regression model is defined as

$$y_i = \beta_0 + \sum_{j=1}^{d} f_j(x_{ji}) + \varepsilon_i, \tag{2.21}$$

where $f_j(\cdot)$ are univariate smooth unknown functions of x_j respectively and β_0 is an intercept. The standard approach to estimate (2.21) is the backfitting algorithm which is described in Hastie and Tibshirani (1990) [2.8]. In order to ensure the functions f_j are uniquely identifiable, the intercept term can be held at $\hat{\beta}_0 = \bar{y}$, the sample mean, and the additional term can be held, $\sum_{j=1}^{d} f_j(x_{ji}) = 0$. An easier way to estimate (2.21) is to use penalized splines. For example, let $d = 2$ and a truncated linear basis should be used. By substituting (2.2) to (2.21), a penalized spline is obtained as follows,

$$y_i = \beta_0 + \beta_{11}x_{1i} + \sum_{k=1}^{K_1} u_{1k}(x_{1i} - \kappa_{1k})_+ + \beta_{21}x_{2i} + \sum_{k=1}^{K_2} u_{2k}(x_{2i} - \kappa_{2k})_+ + \varepsilon_i, \tag{2.22}$$

(2.22) can be fitted by penalized least square or by rewriting it as a linear mixed model with two random effects. The individual components of (2.21) can also be extended to the bivariate smooth functions or one variable with varying coefficients.

2.5 Generalized Smoothing Model

In (2.1) we saw that we could extend the parametric linear regression to nonparametric regression by using the penalized spline method. Similarly, here we relax the linear part of the generalized linear model assumptions and work in a nonparametric framework. Considering that Chapter 5 depends on this model and it is relatively more complex than the aforementioned models, we give more details in

[2.8] See page 91, Hastie and Tibshirani (1990)

this section. We now consider the following structure as an extension of the generalized linear model. We have $y_1, y_2, \ldots, y_n$ independent response observations with means $\mu_1, \mu_2, \ldots, \mu_n$ respectively. The model involves regressors $x_1, x_2, \ldots, x_k$. And y_i are drawn from a member of the exponential family with density expressed in the form,

$$f(y_i; \tilde{\theta}_i, \phi) = exp\{\frac{y\tilde{\theta}_i - b(\tilde{\theta}_i)}{a(\phi)} + c(y_i, \phi)\}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are specific functions. The parameter $\tilde{\theta}_i$ and ϕ are a canonical parameter and a dispersion parameter respectively. The response expectation is linked with canonical parameter by $\mu_i = b'(\tilde{\theta}_i)$. A generalized linear model is identified through a link function $g(\cdot)$,

$$g(\mu_i) = \eta_i \qquad i = 1, 2, \ldots, n$$

where $\eta_i = \mathbb{X}_i'\gamma = \gamma_0 + \sum_{j=1}^{k} \gamma_j x_{ji}$. $g(\cdot)$ is denoted as a canonical link if it is chosen so that $\eta_i = \tilde{\theta}_i$. Here we have the response observations distribution and choose the link function $g(\cdot)$ as a transformation on the observations mean $E(y_i)$. For each response distribution there exits a specific corresponding link function, such as identity link, probit link, and logistic link. The extension of generalized response in a nonparametric way lies in η. If we extend the linear predictor $\eta_i = \mathbb{X}_i'\gamma$ to nonparametric form $\eta_i = m(\mathbb{X}_i)$ and $m(\cdot)$ is a smoothing function, we get the following generalized smoothing model with canonical link,

$$f(y_i, \tilde{\theta}_i, \phi) = exp\{\frac{y\tilde{\theta}_i - b(\tilde{\theta}_i)}{a(\phi)} + c(y_i, \phi)\} \tag{2.23}$$

$$g\{E(y_i)\} = \eta_i \quad i = 1, 2, \ldots, n \tag{2.24}$$

$$\tilde{\theta}_i = \eta_i = m(\mathbb{X}_i) \quad i = 1, 2, \ldots, n \tag{2.25}$$

For the sake of simplicity, we here model the multivariate function $m(\mathbb{X}_i)$ by penalized splines, $m(\mathbb{X}_i) = X_i\beta + Z_i u$. Combing (2.23), we have response density as follows,

$$f(y_i|m(\mathbb{X}_i),\phi) = exp\{\frac{y_i(X_i\beta + Z_iu) - b(X_i\beta + Z_iu)}{a(\phi)} + c(y_i,\phi)\}$$

The likelihood is

$$l(y,\beta,u) = \prod_{i=1}^{n} f(y_i|m(\mathbb{X}_i),\phi)$$

The log-likelihood function is,

$$L = \log l(y,\beta,u) = \sum_{i=1}^{n}\{\frac{y_i(X_i\beta + Z_iu) - b((X_i\beta + Z_iu))}{a(\phi)} + c(y_i,\phi)\} \quad (2.26)$$

Assuming we use the truncated polynomial basis, the parameters β and u can be estimated from the following penalized log-likelihood,

$$L_p = \log l(y,\beta,u) = \sum_{i=1}^{n}\{\frac{y_i(C_i\theta) - b((C_i\theta))}{a(\phi)}\} - \frac{1}{2}\lambda u^T u \quad (2.27)$$

where $c(y_i,\phi)$ is omitted, $X_i\beta + Z_iu = C_i\theta$ and $\theta = (\beta^T, u^T)^T$, $C_i = [X_i, Z_i]$. Maximization of (2.27) can be achieved by penalized iteratively re-weighted least squares (IRLS). The details are as follows,

$$\frac{\partial L_p}{\partial\theta} = \frac{\partial L_p}{\partial\tilde{\theta}}\frac{\partial\tilde{\theta}}{\partial\eta}\frac{\partial\eta}{\partial\theta} \quad (2.28)$$

Since we have the canonical link, then $\eta = \tilde{\theta}$.

$$\frac{\partial L_p}{\partial\theta} = \frac{\partial L_p}{\partial\tilde{\theta}}\frac{\partial\eta}{\partial\theta} \quad (2.29)$$

Recall that $\eta_i = C_i\theta$, which implies that

$$\frac{\partial\eta}{\partial\theta} = C_i^T \quad (2.30)$$

We note that $b'(\tilde{\theta}_i) = \mu_i$

$$\frac{\partial L_p}{\partial \theta} = \sum_{i=1}^{n} C_i^T \{\frac{(y_i - \mu_i)}{a(\phi)}\} - \lambda D\theta \tag{2.31}$$

The score equations are

$$\sum_{i=1}^{n} C_i^T \{\frac{(y_i - \mu_i)}{a(\phi)}\} - \lambda D\theta = 0 \tag{2.32}$$

We first approximate $y_i - \mu_i$ by finding its first-order Taylor series approximation

$$y_i - \mu_i \approx \frac{\partial \mu_i}{\partial \eta_i}(\eta_i^* - \eta_i) \tag{2.33}$$

Combining $\eta_i = \theta_i$ and $Var(\mu) = \frac{\partial \tilde{\theta}_i}{\partial \mu_i}$, we have

$$y_i - \mu_i = Var(\mu)(\eta_i^* - \eta_i) \tag{2.34}$$

Substituting (2.34) into the scoring equations, we rewrite them in matrix form as follows,

$$C^T W(\eta^* - C\theta) = \lambda D\theta$$

Rewriting above equation, the Fisher scoring update can be organized as

$$\theta_{k+1} = (C^T WC + \lambda D)^{-1} C^T Wz$$

where z is a working vector with components $z_i = g'(\mu_i)(y_i - \mu_i) + g(\mu_i)$, and W is the diagonal matrix of working weights $1/[g'(\mu_i)^2 \nu(\mu_i))]$. The updated θ is obtained from weighted penalized smoothing of the working vector z on X. The smoothing parameter λ can be chosen by adjusted GCV or AIC, where the deviance is replaced by Pearson statistic,

$$GCV^p = \frac{n \sum_{i=1}^{n} w_i (y_i - \mu_i)^2}{[n - tr(A)]^2}$$

$$AIC^p = \frac{1}{n} \sum_{i=1}^{n} w_i (y_i - \mu_i)^2 + 2tr(A)/n$$

		mgcv	SemiPar	nlme
Models	Bivariate model	✓	✓	✓
	Varying Coefficients Model	✓	×	✓
	Generalized Smoothing Model	✓	×	×
	Additive Model	✓	✓	✓
Specific Consideration	Correlation structure	✓	×	✓
	REML	✓	✓	✓
	Grouping	✓	✓	✓

Table 2.1: Nonparametric models in different packages. The sign "✓" means that the model or a specific consideration can be realized by the corresponding package while the sign "×" has the opposite meaning.

where w_i is the diagonal elements of W. There are two methods available for selecting the smoothing parameters for the generalized smoothing model. One approach is to select smoothing parameters by minimizing the above criteria in each iteration. The other approach is to iterate to convergence for the given smoothing parameters and find those that minimize the criteria.

2.6 Computation Packages

There are many software packages in current use in addition to the popular *R*, such as *Matlab SAS*. The sample codes for both software can be found in Ruppert et al. (2003). R is used here as an example. Hence the descriptions and programs in Appendix D are intended to show how the common idea of penalized spline is implemented. Table 2.1 is a summary of the use of the packages for models discussed in this thesis. More details can be seen in corresponding reference manuals, which can be downloaded from,

http://cran.r-project.org/web/packages/SemiPar/SemiPar.pdf and

http://cran.r-project.org/web/packages/mgcv/mgcv.pdf.

An introduction to applying mixed model package *nlme* for penalized spline can be found in Ngo and Wand (2004). The generalized mixed model can be applied by the command *glmmPQL* in Package *MASS*.

Chapter 3

Impact of Different Interest Rates on the Prices of MBS

3.1 Introduction to Mortgage-Backed Securities

3.1.1 Overview of MBS Market

Mortgage-backed securities (MBS) are a type of fixed income investment, which are collateralized by residential or commercial mortgage loans. Focus on MBS are important for many reasons. First, mortgage debt contributes to a significant part of the U.S economy, which is illustrated in Figure 3.1. By the first quarter of 2007, home mortgage accounts for 35% of the total $29255 billion outstanding debt in the U.S non-financial sector.[3.1] Second, mortgage-related bonds[3.2] account for the largest percentage of the whole U.S bond market by the first quarter of 2007[3.3]. As shown in Figure 3.2, mortgage-related bonds and corporate bonds together account for 48% of the outstanding debt in the U.S. bond market. Finally, MBS account for one of the largest parts of the securitization market.

[3.1] Source: Federal Reserve, Report Z.1, table D.3

[3.2] Includes GNMA, FNMA, and FHLMC mortgage-backed securities and CMOs and private-label MBS/CMOs

[3.3] Source: Securities Industry and Financial Markets Association

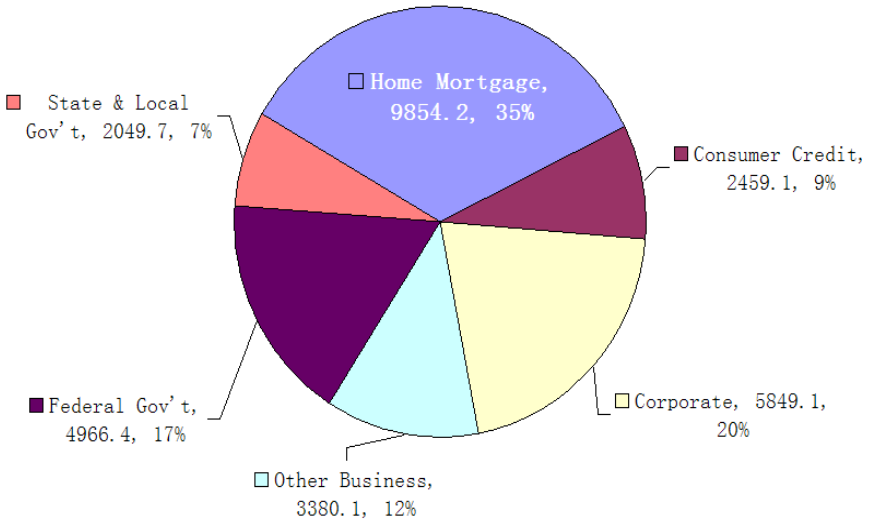


Figure 3.1: Debt Outstanding of Domestic Nonfinancial Sectors by the first quarter 2007

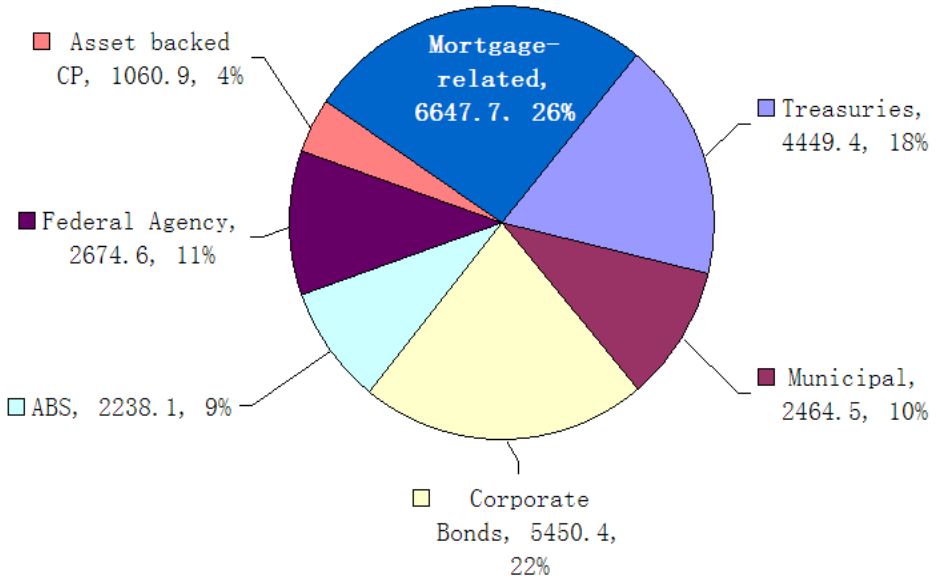


Figure 3.2: Debt outstanding of U.S. Debt Capital Markets by the first quarter 2007

3.1.2 Residential Mortgage Loans Basics

The residential mortgage loans in the U.S. can be divided into categories in two different ways. Firstly, conventional and government loans. Any mortgage loan other than an FHA[3.4], VA[3.5] or an RHS[3.6] loan is a conventional one. Secondly, fixed rate loans, adjustable rate loans and hybrid loans. With fixed rate mortgage (FRM) loan the interest rate and monthly payments remain fixed for the life of the loan. The most common mortgage terms are 30 and 15 years. The borrower has the right to prepay the loan at any time. Another loan is called an adjustable-rate mortgage loan or ARM. The contract rate is adjusted periodically, either annually or semiannually, based on the changes in a predetermined index, such as London Inter Bank Offering Rates (LIBOR), Treasury Bill (T-Bill), 12-Month Treasury Average (MTA or MAT) or Cost of Savings Index (COSI) e.t.c. ARMs typically offer low initial interest rates that last until the first adjustment. Most ARMs avoid enormous increases in monthly payments by a lifetime cap or a periodic cap, which limits the interest rate increase over life and at one time respectively. Some types of ARMs also offer payment caps to limit the amount that the monthly payment can increase. Hybrid loan is a combination of fixed and ARM loans. Some hybrid loans provide a fixed rate for the first several years before periodical adjustments like fixed-period ARMs while the other comes with an option to convert them to a fixed-rate mortgage at designated times after periodical adjustment like convertible ARMs.

3.1.3 Mortgage Backed Securities Basics

Mortgage backed securities are bonds backed by mortgage loans, which are products of securitization. The most common type of MBS is called pass through, representing the participation certificates. The originator selects the mortgage loans to create a pool. An investor who owns MBS receives interest and principal from the servicer. The servicer collects the payment from mortgage borrowers and distributes them to investors.

Most MBS are issued or guaranteed by one of the following three agencies, Government National Mortgage Association (Ginnie Mae), a U.S. government agency,

[3.4]The Federal Housing Administration
[3.5]U.S. Dept. of Veterans Affairs
[3.6]The Rural Housing Service (RHS) of the U.S. Dept. of Agriculture

or the Federal National Mortgage Association (Fannie Mae) and the Federal Home Loan Mortgage Corporation (Freddie Mac), U.S. government-sponsored enterprises. Each agency provide credit guarantee for the corresponding MBS. Ginnie Mae's guarantee is backed by federal government. The guarantees of the other two agencies do not have federal backing, but are regarded by market participants as possessing extremely high credit quality.

Ginnie Mae is the trade name of the Government National Mortgage Association (GNMA). Ginnie Mae does not issue any securities. A mortgage banker first generate a pool of residential mortgage loans, then applies to Ginnie Mae for a guarantee commitment. Once Ginnie Mae approves the application, the originator receives a pool number for the mortgage pool and then issues mortgage pass through certificates. The issuer can hold the mortgage pass through certificates in its portfolio or sell the certificates to investors or to dealers. Subsequently, the servicer is responsible for serving the loans in the pool. Ginnie Mae guarantees full and timely payment of principle and interest on its MBS. It offers two programs called Ginnie Mae I and Ginnie Mae II. All mortgage loans in Ginnie Mae I program must be of the same type, have the same interest rates and be issued by the same issuer. Hence, all loans in Ginnie Mae I program are approximately homogeneous. The Ginnie Mae II program allows a wide range of coupons. The Ginnie Mae I program does not accept ARMs, but the Ginnie Mae II program does.

Fannie Mae and Freddie Mac are "government-sponsored enterprises (GSEs). Although they are privately owned, they receive support from the Federal Government, and assume some public responsibilities which is to make sure mortgage money is available for people in communities in America. Fannie Mae and Freddie Mac accept both government mortgage loans and conventional mortgage loans. However, their main emphasis is on conventional mortgage loans. They do not lend money directly to home buyers, instead, they operate in the secondary mortgage market to make sure the lenders do not run out of mortgage funds. Their business focuses on two perspectives. First, they purchase mortgages from lenders and hold them in their portfolio, keeping money flowing to mortgage lenders. Second, they issue mortgage-backed securities (MBS) in exchange for pools of mortgages from lenders. For lenders, these MBS are more liquid than mortgages.

3.1.4 Risks Related to MBSs

Interest Rate Risk

Like any other fixed-income instrument, mortgage-backed securities bear exposure to interest rate risk. If interest rates decrease, the payment from mortgage borrowers may accelerate, which in turn contracts average life. Interest payments will only be received over a short period of time and investors have to reinvest their interest income and any return of principal at lower prevailing rates. Conversely, if interest rates increase, return of payment can decelerate, causing the security's average life to extend. The rising interest rates delay the return of principal to their investors and cause them to miss the opportunity to reinvest at higher yields. In either case, changes in the level of interest rates can directly affect a mortgage-backed security's market value and total return. This topic is partly investigated in the remaining sections in this chapter.

Prepayment Risk

As previously discussed, a major risk in an investment in MBS is prepayment risk which is caused by the changes in the prepayment speeds of the underlying mortgages. All FHA and VA mortgage loans can be prepaid at any time without any penalty. An increase in prepayment speeds results in a faster decline of the principle than what may be expected and a contraction to the average life. Conversely, a decrease in prepayments will result in a slow down in principal returns and an extension to the average life. Market participants have developed prepayment models to evaluate prepayment risks. The burnout effect caused by the prepayment behavior is discussed in next chapter.

Credit Risk

Credit risk refers to the fact that the mortgage borrowers may not make timely payments or may default on their loans. Credit risk is thought to be more significant in private-label MBS than in GSEs or Ginnie Mae MBS because the GSEs and Ginnie Mae guarantee the timely payment of principal and interest on the MBS. As discussed above, the Ginnie Mae guarantee is backed by the full faith and credit of the United States. Fannie Mae and Freddie Mac guarantee also has high credit

quality. Generally, the credit risk of most mortgage-backed securities carry bond insurance that guarantees minimal payments of interest and principal to investors. The guarantee of principal and timely payment means that the credit risk can be removed from the consideration. With a defaulted mortgage, the payments will continue until the principal amount is repaid by the guarantor. For example, if the securities issuer defaults on the monthly payment, GNMA is responsible for the full and timely payment of principle and interest. The investor also receives additional amounts for settlement on those loans in the pool which have been foreclosed. Thus, the insured default may appear to the lender as a prepayment, but the incidence of these prepayment is nonetheless affected by the factors different from prepayment, such as house prices.

3.2 Motivation of Nonparametric Modelling

Mortgage backed securities (MBS) are financial assets backed by a pool grouped by a large number of mortgages. Payments of principal and interest that mortgagors make every month are passed through to investors. In this chapter, one kind of MBS, namely, GNMAs, will be investigated, with the special feature that the timely payment of principal and interest is guaranteed by the Government National Mortgage Association (GNMA). In terms of prepayment modelling, the approaches to pricing MBS are divided into two categories, namely, the structural approach and the reduced form approach respectively. The structural approach models prepayment and default behavior as exercising a call option or a put option. Dunn and McConnell (1981a) and Dunn and McConnell (1981b) developed an optimal prepayment model. Timmis (1985), Dunn and Spatt (1986) and Johnston and Drunen (1988) extend this optimal strategy by also considering transaction cost. Stanton (1995) introduces heterogeneity of the transaction costs into the rational prepayment model discussed above and applies discrete prepayment. Kau et al. (1992), Kau et al. (1995), and Kau and Keenan (1995) explore a two-factor option-pricing model. Deng et al. (2000) apply a competing risk model to model prepayment, default and heterogeneity of borrowers using loan-level data. Downing et al. (2005) introduce the impact of house prices into a valuation model handling both prepayment and default using pools data. The reduced form approach model the prepayment or default as a function of selected predictors without any theoreti-

cal restrictions. Schwartz and Torous (1989) apply methods in survival analysis to prepayment modelling. Although parametric models are popular with academics, nonparametric techniques are also applied to problems related to MBS. Boudoukh et al. (1995), Boudoukh et al. (1997), LaCoure-Little et al. (1999) and Maxam and LaCour-Little (2001) demonstrate the application of kernel based nonparametric approach to prepayment modelling. To avoid the curse of dimensionality suffered by the kernel approach, Jegadrrsh and Ju (2000) model prepayment rate using another nonparametric approach, generalized additive model (GAM). We follow this nonparametric trend to explore the impact of different interest rates by using penalized spline smoothing, as recent powerful smoothing techniques.

Most of the above mentioned papers assume that short term interest rates and long term interest rates play important roles in MBS pricing. The reason why we are interested in the impact of interest rates on prices is related to hedging issues. Hedging the interest rate risk of MBS is to construct a portfolio including MBS and make this portfolio insensitive to the changes of interest rates. It will be easier to find such portfolio if we have a better understanding of the following questions. Is the effect of the long term interest rate on the price the same as that of the short term interest rate ? Moreover, how do the effects change over time ? We usually get the answers adapting a structural method, namely, first get the price of MBS as a numerical solution of some partial differential equation with boundary conditions, [3.7]then analyze the plot of price against interest rates or calculate the effective duration as the sensitivity of the price to changes in interest rates. However, this method not only depends on the choice of interest rate models and prepayment models, but also can not include non financial factors. Moreover, the pool-level focus is complicated for portfolio management. We usually can find a large number of GNMA pools included in the portfolio positions of many funds, such as, Pioneer Government Income Fund or HighMark Balanced Fund. Thus, a method that hedges the homogeneous parts of the whole GNMA portfolio instead of pool by pool against interest rates risk, will be more efficient. We propose a new strategy to hedge interest rates risk in this way. Firstly, we investigate the impact of different interest rates on prices of MBS, then turn to the hedging issue.

Figure 3.3 shows the scatter plots of some MBS prices against interest rates. The visual impression differs from both the theoretic indication that MBSs in the

[3.7]See Kau and Keenan (1995), Gaussel and Tamine (2004).

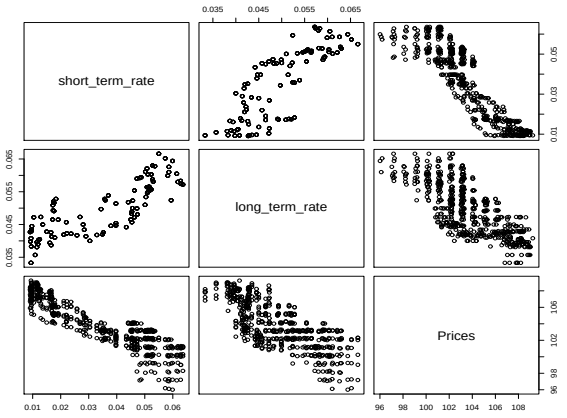


Figure 3.3: Scatter plots of prices against long term interest rates and short term interest rates for GNMA issued 1992 with 15 years

end phase should be insensitive to changes in the long term interest rates and the simulation result in Schwartz and Torous (1989), Stanton (1995), Jegadrrsh and Ju (2000), where the prices of MBS were found to significantly decrease at some critical point as the long term interest rate decreased. The difference highlights the demand for a model that is more capable of assessing the relationships among variables. For this purpose we use nonparametric method. We also know that the long term interest rate and the short term interest rate have influences on the prices of MBS, furthermore, these influences are nonlinear from an economic point of view. Meanwhile, if we observe the market, what we could potentially get is panel data, namely, the prices of MBS with different coupons and different maturities. To cope with problems in this setting, we prefer the mixed additive models, which are able to find the nonlinear relationship and account for heterogeneity by using grouping factors simultaneously. Pinheiro and Bates (2000) explain the details and uses of mixed models.

In contrast to other nonparametric works the P-spline approach has the following advantages. Firstly, we would avoid the extrapolation problem or curse of dimensionality. Secondly, this approach models the panel structure of the market MBS prices data, which is consistent with the portfolio management. Thirdly, and most importantly, the influences of the long term interest rate and the short term interest rate could be illustrated respectively, and hence we could identify the evolution

of the impact of interest rates over time. Finally, we also could directly estimate derivatives for hedging purpose.

3.3 Nonparametric Modelling

Nonparametric methods have become popular in the last two decades, Kauermann (2006) provides a summary of the main nonparametric models. Non-parametric methods also have been used to estimate MBS prices in Boudoukh et al. (1995) and Boudoukh et al. (1997) or prepayment rates in LaCoure-Little et al. (1999), Jegadrrsh and Ju (2000), Maxam and LaCour-Little (2001). Boudoukh et al. (1997) applies a nonparametric technique named multivariate density estimation (MDE) to estimate MBS TBA prices. Due to the curse of dimensionality of MDE they use only two variables, the level of interest rates and the slope of the term structure. Among these nonparametric models we will use the additive mixed model discussed in Section 2.4 for the purpose of identifying the impact of different interest rates. An example of an additive mixed model with single grouping factor has a structure like the following:

$$y_{ij} = U_i + f_1(x_{1j}) + f_2(x_{2j}) + \varepsilon_{ij} \tag{3.1}$$

where i is a group index, y_{ij} is a univariate response for group i; the f_1 and f_2 are smooth functions of covariates x_1 and x_2 respectively; U_i is the specific effect to group i ; ε is a residual error vector. In order to fit the nonparametric model (3.1) we first represent the unknown functions as penalized spline, then the estimation of additive model (3.1) is transformed to estimate a mixed model, which has previously been discussed. The only difference in estimation is to account for the grouping factor by adding more columns to matrix Z in (3.3), for details see Appendix A.

3.4 Data and Empirical Modelling

3.4.1 Data

Our empirical analysis focuses on the functional relationship between the prices of different mortgages pools and interest rates. To outline the evolution of this rela-

Coupon Rate	6.95%	7%	7.5%	8%	8.5%	9%	9.5%	10%
Min	94.15	95.26	97.31	100	101.2	102.1	102.3	104.1
Mean.	102.5	103.2	104.6	105.9	106.9	108	108.3	109.9
Max.	106.7	107.5	108.6	109.8	110.5	112.1	111.8	112.9

(a) Issued in 1992 with 30 years Maturity, Observed in July 98-March 2006

Coupon Rate	9.5%	10%	10.5%	11%	11.5%	12%	12.5%	13%
Min	102.3	103.3	105.1	106.1	107.1	107.2	109.2	109.1
Mean.	108.4	109.3	110.1	111	112.2	113.1	114.4	114.7
Max.	111.5	111.9	112.1	113.4	114.7	115.1	116.2	117.2

(b) Issued in 1983 with 30 years Maturity, Observed in August 97-March 2006

Coupon Rate	6.5%	7%	7.5%	8%	8.5%	9%
Min	96.01	97.22	99.09	100	100.1	100.8
Mean.	103.00	103.6	104.2	104.1	104.1	103.6
Max.	109.2	108.9	109	108.5	107.7	107.3

(c) Issued in 1992 with 15 years Maturity, Observed in August 97-March 2006

Table 3.1: Descriptive statistics of prices of different GNMAs

tionship, we use three groups of GNMA II pools. Moreover, each group includes pools issued in the same year and with the same maturity but with different pass through coupon rates. We regard the functional relationship as a common characteristic of the corresponding group. The pools data are collected from Reuters 3000 Xtra[3.8]. It consists of 2038 monthly prices over the period from August 1997 to March 2006. As mentioned above, the data are classified into three groups, namely, 6 GNMAs with 15 years maturity issued in 1992, 8 GNMAs with 30 years maturity issued in 1992, 8 GNMAs with 30 years maturity issued in 1983. Table 3.1 shows more statistical details.

Following Schwartz and Torous (1992), Boudoukh et al. (1997) [3.9] and Kariya

[3.8] http://about.reuters.com/productinfo/3000xtra/description.aspx

[3.9] Boudoukh et al. (1997) illustrate that the yield of 10-year Treasury note is closely correlated with the average 30-year mortgage rate.

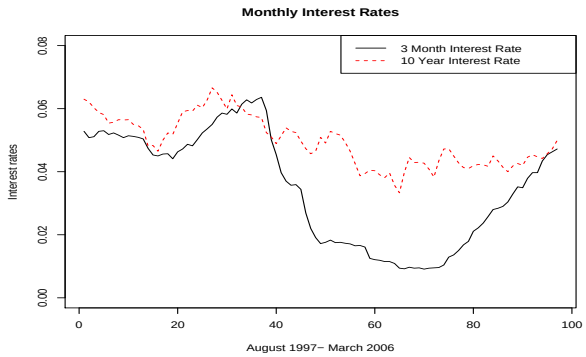


Figure 3.4: Observed monthly short term interest rate and long term interest rate for period August 1997 - March 2006

et al. (2002), we also use the yield of 10-year Treasury note as a proxy of the mortgagor's incentive to prepay the mortgage. The yield of 3-month Treasury bill represents the overall level of interest rates and plays a role of discounting factor in pricing. These interest rates represent short term and long term interest rate respectively. There are two arguments in support of our choice. On the one hand, theory and actual prepayment experience indicates that the spread between current mortgage rate and the mortgage contract rate is inversely related to the prepayment rate. On the other hand, the two factors usually account for 90-95% of the observed variability of the yield curve.[3.10]

[3.10] See Rebonato (1998), P61 .

Figure 3.4 illustrates the trend of 10 year T-note yield and 3 month T-bill yield during the period 1997 to 2006. The high volatility in 3 month interest rate path allows us to verify our assumption on its relationship with the MBS prices. Compared with the strong fluctuation in the 3 month interest rate the 10 year interest rate demonstrates a moderate decline. In this period the dynamics of the 3 month interest rate acts as a mean reverting to the 10 year interest rate.

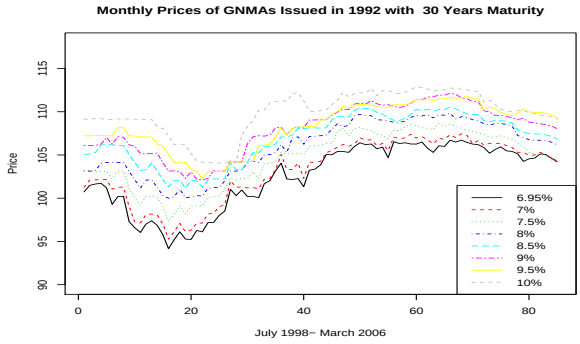
Figure 3.5 shows three groups of the prices of different GNMAs for successive months. Figure 3.5(a), Figure 3.5(b) and Figure 3.5(c) represent different issue dates respectively. The responses of prices in each plot are apparently similar except that they are of different magnitudes. However, there is also some evidence for differences among the curves.

3.4.2 Empirical Modelling

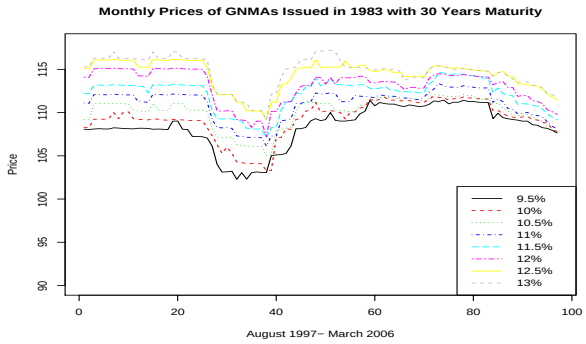
Inspection of Figure 3.5 indicates that the evolutions of the curves in each plot have almost the same direction but with different magnitudes of variability at a point in time. Therefore, we would expect that distinguishing GNMAs from each other by magnitudes would be beneficial. Normally we regard coupon rates as a important factor which determines the effect of the interest rates on prices. Due to this, the different magnitudes of the responses of prices to interest rates can be modelled by a random intercept controlling for individual heterogeneity mainly related to coupon rates. The phenomenon of the same direction of the curves can be characterized by smooth functions representing the nonlinear relationship between prices and interest rates. To cope with this panel data, it is best to select additive mixed models containing random effects and fixed effects in the context of analyzing a group of GNMAs with different coupons simultaneously, which are better able to identify and measure the impacts of interest rates that are not detectable in pure time series data of one GNMA.

The data are categorized according to time to maturity: 1992 with 30 years maturity, 1992 with 15 years maturity, 1983 with 30 years maturity. For each group we fit the model grouped by coupon rates using an additive structure, and then show how the impact of interest rates on GNMAs change over time by analyzing three fitting results.

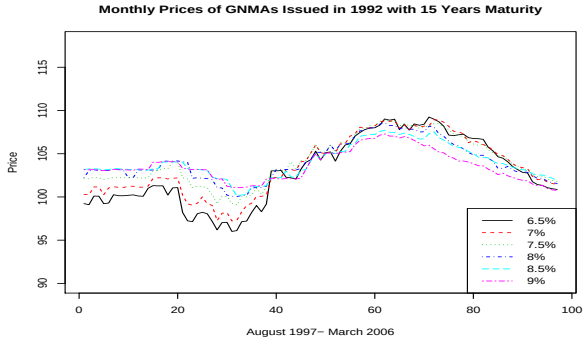
As other option-free bonds, the prices of MBS can be calculated through the



(a)



(b)



(c)

Figure 3.5: Observed monthly prices of GNMAs with different coupon rates and issue dates for period August 1997 - March 2006

net present value of a set of future cash flows. However, due to the prepayment behavior the cash flows of MBS are uncertain. The relationship between MBS prices and interest rates is nonlinear. To address the main question of interest, we have the following model for log prices of GNMAs,

$$p_{ij} = c_i + f(s_{ij}) + g(l_{ij}) + \varepsilon_{ij} \tag{3.2}$$

where p_{ij} denotes the logarithmized price of GNMA$_i$, i=1, ..., n, on date j, $j = 1, ..., T$. c_i i.i.d.$N(0, \sigma_c^2)$ is the random effect induced by the coupon rate i. f and g are smooth functions of short term interest rate and long term interest rate respectively. Financial data are taken to be serially correlated and correlated errors are usually considered by mixed model software, hence, for the residual structure we assume an AR(1) process of the form $\varepsilon_{ij} = \rho \cdot \varepsilon_{i,j-1} + \xi_{i,j}$, $\xi_{i,j}$ i.i.d. $N(0, \sigma_\xi^2)$. $f(s_{ij})$, $g(l_{ij})$ can be thought of as the systematic part of the price, whereas c_i can be regarded as heterogeneity between pools and $\varepsilon_{i,j}$ represents heterogeneity within a pool.

Fitting model (3.2) via penalized splines is equivalent to looking for a mixed model and estimating the variance components for random intercept and the amount of smoothing for f and g. Hence we rewrite additive mixed model (3.2) in matrix form as follows,

$$\tilde{P} = X\beta + Zu + \varepsilon, \tag{3.3}$$

The details of (3.3) can be found in Appendix. Table. 3.2 shows the estimation results of model (3.2). Once we get $\hat{\beta}$ and $\hat{u}$ the nonlinear relationship between prices and interest rates can be illustrated by the estimated function $\hat{f}$ and $\hat{g}$ with intercepts constrained to zero.

$$\hat{f}(s_j) = \hat{\beta_{11}} s_j + \hat{\beta_{12}} s_j^2 + \sum_{k=1}^{k_s} \hat{u}_{k_s} (s_j - \kappa_{k_s})_+^2 \tag{3.4}$$

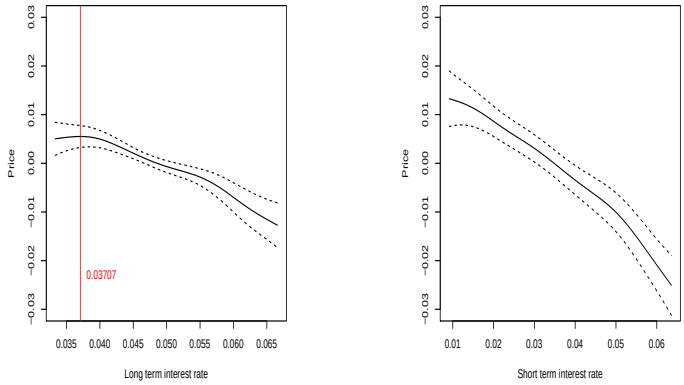
$$\hat{g}(l_j) = \hat{\beta_{21}} l_j + \hat{\beta_{22}} l_j^2 + \sum_{k=1}^{k_l} \hat{u}_{k_l} (l_j - \kappa_{k_l})_+^2 \tag{3.5}$$

3.5 Impact of Different Interest Rates

Figure 3.6, Figure 3.8 and Figure 3.10 show the curves representing the relationship between interest rates and prices in different maturities. The GNMAs used to generate Figure 3.6, Figure 3.8 and Figure 3.10 were issued in 1992, 1983, 1992 and with 30 years, 30 years and 15 years maturity respectively. To simplify this, we assume that they represent the early period, middle period and end period of GNMAs. The log prices are constrained to have a mean of zero. The right plots (a) represent the effect of the long term interest rates on the prices and the left plots (b) illustrate how short term interest rates impact the prices. By comparing these curves we can achieve a profound understanding of the impact of different interest rates.

Firstly, there is evident difference between the shapes representing long and short term interest rates. In right plots the prices increase as the short term interest rate decrease, which is verified by the negative first derivatives in Figure 3.7 (d), Figure 3.9 (d) and Figure 3.11 (d). However, the curves in the left plots illustrate a falling trend when the long term interest rates reach certain points. Analogously, it is also verified by the phenomenon that the first derivatives become positive for some long term interest rates. We can find these values by calculating the points where the first derivatives become significantly positive as the long term interest rates decrease. These values, 0.03707, 0.06055, 0.03825, are illustrated by vertical lines in plots (a) in Figure 3.6, Figure 3.8 and Figure 3.10. This falling trend reflects the well known negative convexity of MBS, namely, due to the refinancing incentive the mortgagors begin to prepay as the long term interest rate decreases and so cause falling prices of MBS.

Secondly, the points where the long term interest rate shows the negative convexity are related to both coupon rates and the maturity period. In contrast to the values, 0.03707 and 0.03825, in Figure 3.6 (a) and Figure 3.10 (a), the turning point in Figure 3.8 (a) is 0.06055, which is consistent to the economic intuition that the MBSs issued in 1983 with higher coupon rates should show prepayment at larger long term interest rates than those issued in 1992 with relatively smaller coupon rates. To identify the role of coupon rate more precisely, we calculate the differences between the turning points and average coupon rates of the three groups to represent the incentive of prepayment of the corresponding group, which is 0.0459925 for early period with average coupon rate 0.0830625, 0.0519 for middle period with

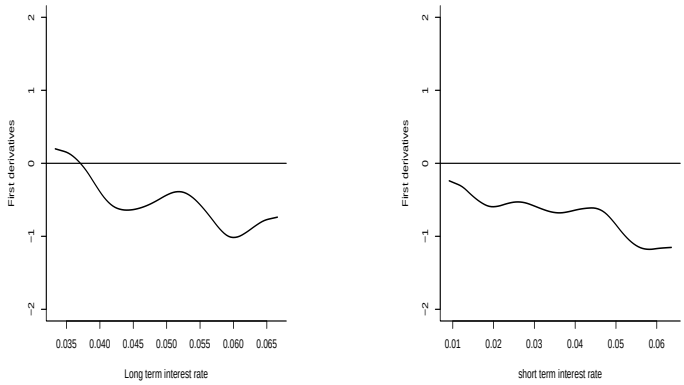


(a) (b)

Figure 3.6: Fitted short term interest rate and long term rate effects on the prices of GNMAs issued in 1992 with 30 years maturity

average coupon rate 0.1125 and 0.03925 for end period with average coupon rate 0.0775. The larger difference values correspond to the higher coupon rates of GNMAs. Figure 3.6 (a) shows that the long term interest rate has almost no impact on the prices of GNMAs issued in 1983 in the range of its lower values, which is consistent to the burnout effect of the prepayment behavior. In other words, prepayment is decreasing over time and does not appear even when the spread between mortgage contract rate and refinance rate are large.

Finally, the first derivatives, which measure the sensitivity of prices to different interest rates also show different dynamic behavior in each period. Based on the estimated derivatives at real interest rates, Table 3.3 shows that the absolute values of the derivatives of short term interest rates are not only larger than those of long term interest rates, but are also more variable. The combination of these plots indicates that the impact of long term interest rate is more evident in the early period than in the end stage. Conversely, the impact of short term interest rates in end stage is larger than the impact in early period.



(c) (d)

Figure 3.7: Fitted first derivatives corresponding to Figure 3.6

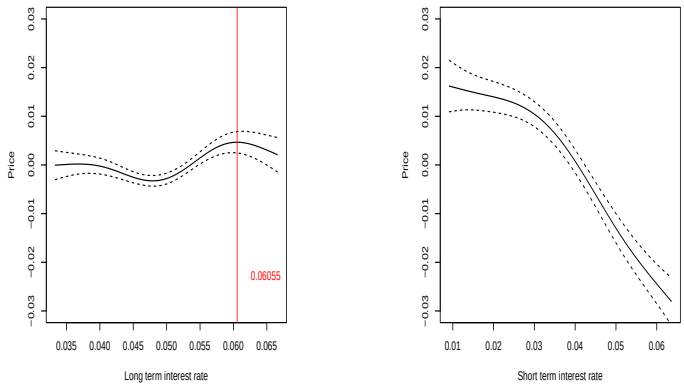
3.6 Hedging MBS Portfolio against Interest Rates Risk

One advantage of our model is its convenience to calculate the derivatives for hedging. Ruppert et al. (2003) point out that in mixed model context the optimal smoothing parameter depends not on the order of derivatives to be estimated, but only on the variance components. Furthermore, its value for estimation of the first derivative or the second derivative is close to that for the estimation of the function.

In this section we first show how to predict the price changes of a portfolio consisting of one of the three GNMA groups discussed above in nonparametric framework given small changes in interest rates, and then discuss some aspects of hedging such a portfolio against interest rate changes.

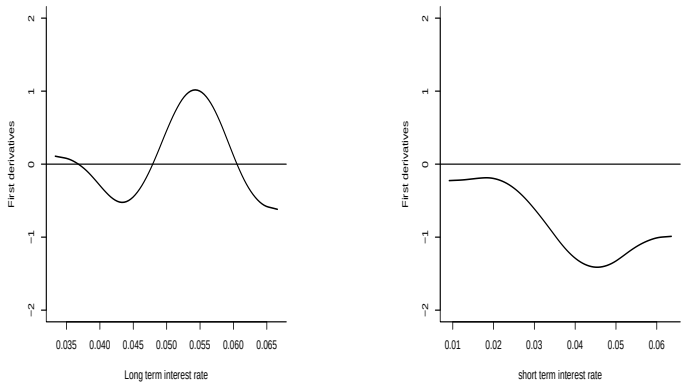
3.6.1 Predict the Price Change due to Changes in Interest Rates

Since not all yield curve shifts are parallel moves, we assume that MBS prices are affected by both short term and long term interest rates. We extend and rearrange the Taylor series of prices around the current interest rates to first order without



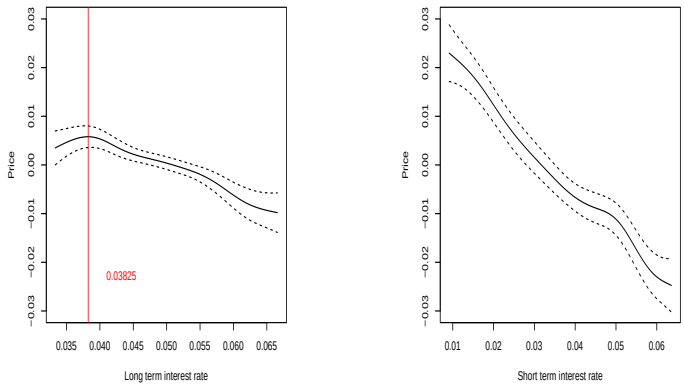
(a) (b)

Figure 3.8: Fitted short term interest rate and long term rate effects on the prices of GNMAs issued 1983 with 30 years maturity



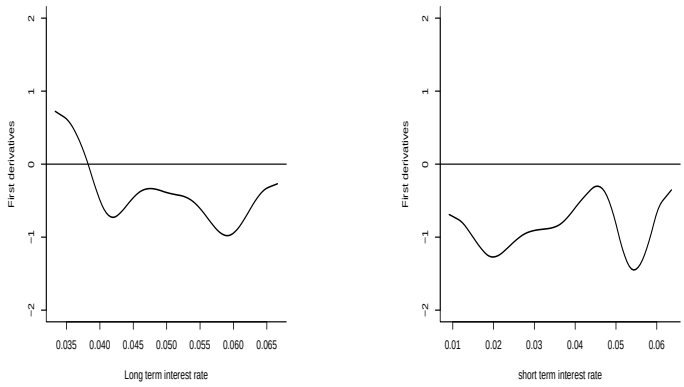
(c) (d)

Figure 3.9: Fitted first derivatives corresponding to Figure 3.8



(a) (b)

Figure 3.10: Fitted short term interest rate and long term rate effects on the prices of GNMAs issued 1992 with 15 years maturity



(c) (d)

Figure 3.11: Fitted first derivatives corresponding to Figure 3.10

Non-linear component	edf	F	p-value
Long term	4.439	7.303	1.85e-08
Short term	3.785	11.697	4.37e-14
n	680	Loglik	2545.215

(a) GNMAs issued in 1992 with 30 years Maturity over the period of July 98-March 2006

Non-linear component	edf	F	p-value
Long term	3.869	11.76	2.8e-09
Short term	3.529	40.05	<2e-16
n	776	Loglik	3030.16

(b) GNMAs issued in 1983 with 30 years Maturity over the period of August 97-March 2006

Non-linear component	edf	F	p-value
Long term	5.38	7.095	3.8e-08
Short term	5.667	18.708	<2e-16
n	582	Loglik	2293.854

(c) GNMAs issued in 1992 with 15 years Maturity over the period of August 97-March 2006

Table 3.2: Estimation results for three groups

		9230	8330	9215
short term	mean	-0.673	-0.796	-0.854
	variance	0.079	0.249	0.095
long term	mean	-0.592	0.062	-0.538
	variance	0.0552	0.286	0.075

Table 3.3: Mean and variance of estimated first derivatives

considering the higher order terms,

$$\Delta P \approx \frac{\partial P}{\partial r_s}\Delta r_s + \frac{\partial P}{\partial r_l}\Delta r_l \tag{3.6}$$

where ΔP is the change of MBS price due to changes in interest rates, Δr_s and Δr_l are changes of short and long term interest rates. $\frac{\partial P}{\partial r_s}$ and $\frac{\partial P}{\partial r_l}$ are sensitivities of corresponding prices to short and long term interest rates. Since we allow the log price in the previous estimation step, the estimated derivatives $\frac{\hat{\partial P}}{\partial r_s}$ and $\frac{\hat{\partial P}}{\partial r_s}$ can be recovered from the following equations,

$$\frac{\hat{\partial P}}{\partial r_s} = P\frac{\hat{\partial f}}{\partial r_s} \qquad\qquad \frac{\hat{\partial P}}{\partial r_l} = P\frac{\hat{\partial g}}{\partial r_l} \tag{3.7}$$

where $\frac{\hat{\partial f}}{\partial r_s}$ and $\frac{\hat{\partial g}}{\partial r_s}$ are the estimated first derivatives by P-spline approach, shown in Figure 9. We get the following equation to approximate price change due to changes in interest rates by substituting (3.7) into (3.6),

$$\Delta P \approx P(\frac{\hat{\partial f}}{\partial r_s}\Delta r_s + \frac{\hat{\partial g}}{\partial r_l}\Delta r_l) \tag{3.8}$$

In empirical analysis it is quite probable that the future interest rates are out of the range of sample. In this case, we use Taylor expansion from the boundary of the smoothing function to get the estimated derivatives. For example, the estimated first derivative $\hat{d}(r_{s1})$ at r_{s1} is as follows,

$$\hat{d}(r_{s1}) \approx d(r_{s0}) + d'(r_{s0})(r_{s1} - r_{s0}) \tag{3.9}$$

where $d(\cdot)$ is the derivative function of short term interest rate, $d(r_{s0})$ is the derivative at the boundary r_{s0} and $d'(r_{s0})$ is the first derivative of $d(\cdot)$ at r_{s0}. We use a portfolio consisting of GNMAs issued in 1992 with 15 years maturity as an example to illustrate how to predict the price changes by (3.8). To check the out-of-sample performance of our model, we take the observations of last 17 months as a test sample. Recall that we already have the estimated function form of $\frac{\partial f}{\partial r_s}$ and $\frac{\partial g}{\partial r_l}$ in Figure 3.11. Combining (3.9), the first derivatives of short and long term interest rates from August 1997 to March 2006 can be obtained. Thus, we can figure out the price changes due to interest rates using (3.8) and compare it with the empiri-

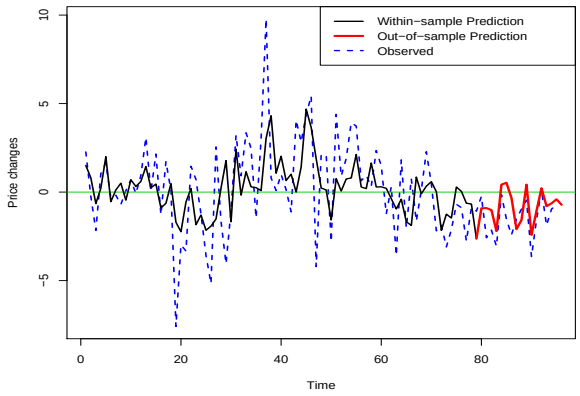


Figure 3.12: Observed price changes and estimated price changes due to changes in interest rates of the portfolio consisting of GNMAs issued in 1992 with 15 years maturity over the period of August 97-March 2006

cal price changes. Figure 3.12 shows the observed price changes and the estimated prices changes due to changes in interest rates for both within-sample and out-of-sample observations. For most of the observations the price changes due to changes in interest rates can explain a substantial part of the observed prices changes. As shown in Figure 3.12, our approach works effectively in out-of-sample test. A more prudent hedging strategy may call for higher order derivatives, which can be similarly obtained from the estimation result as first derivatives.

3.6.2 Hedging Positions

As discussed above, the additive mixed model captures the mean characteristics of a portfolio. The estimated derivatives can not only be applied to the prediction of price changes due to interest rates but also to hedge interest rates risk. These derivatives play a similar role to the conventional duration and convexity. Since we consider both short term and long term interest rates here, it would be best to use both short and long term interest rate futures, for instance, 90-day Treasury bill futures and 10-year Treasury note futures to offset the risks. It is assumed that the changes of future prices due to changes in interest rates can be approximated as follows,

$$\Delta F_s \approx \frac{\partial F_s}{\partial r_s}\Delta r_s + \frac{\partial F_s}{\partial r_l}\Delta r_l$$

$$\Delta F_l \approx \frac{\partial F_l}{\partial r_s}\Delta r_s + \frac{\partial F_l}{\partial r_l}\Delta r_l$$

where ΔF_s and ΔF_l are changes of short term interest rate futures price and long term interest rate futures price respectively. Δr_s and Δr_l are changes of short and long term interest rates. $\frac{\partial F_s}{\partial r_s}$, $\frac{\partial F_s}{\partial r_l}$, $\frac{\partial F_l}{\partial r_s}$ and $\frac{\partial F_l}{\partial r_l}$ are sensitivities of corresponding prices to short and long term interest rates. We assume that the future positions w_s and w_l exactly offset changes in the price of MBS in (3.6) being hedged,

$$\Delta P + w_s \Delta F_s + w_l \Delta F_l = 0$$

Matching Δr_s and Δr_l we produce,

$$\frac{\partial P}{\partial r_s} + w_s \frac{\partial F_s}{\partial r_s} + w_l \frac{\partial F_l}{\partial r_s} = 0$$

$$\frac{\partial P}{\partial r_l} + w_s \frac{\partial F_s}{\partial r_l} + w_l \frac{\partial F_l}{\partial r_l} = 0$$

then future positions w_s and w_l can be calculated by solving the above equations,

$$w_s = \frac{\frac{\partial P}{\partial r_s}\frac{\partial F_l}{\partial r_l} - \frac{\partial F_l}{\partial r_s}\frac{\partial P}{\partial r_l}}{\frac{\partial F_l}{\partial r_s}\frac{\partial F_s}{\partial r_l} - \frac{\partial F_s}{\partial r_s}\frac{\partial F_l}{\partial r_l}} \tag{3.10}$$

$$w_l = \frac{\frac{\partial F_s}{\partial r_s}\frac{\partial P}{\partial r_l} - \frac{\partial P}{\partial r_s}\frac{\partial F_s}{\partial r_l}}{\frac{\partial F_l}{\partial r_s}\frac{\partial F_s}{\partial r_l} - \frac{\partial F_s}{\partial r_s}\frac{\partial F_l}{\partial r_l}} \tag{3.11}$$

3.7 Conclusions and Future Extensions

In this chapter we applied P-spline approach to investigate the relationship between the prices of MBS and interest rates. We find that this technique can easily produce the nonlinear functions of the short term interest and the long term interest rate respectively. Due to the flexibility from its nonparametric nature, penalized spline method also enjoys the ability to capture the evolution of the changing impact of

interest rates on MBS prices along time to maturity when we implement additive mixed models to the data of different maturity. Based on the additive models and their convenience to estimate the derivatives we also propose an approach to hedge the interest rates risk of the MBS portfolios consisting of one of the group discussed above.

Future work might begin by using more data from different pools such as TBA data set. Another extension of this empirical analysis is to apply the method to GNMA securities representing different risk characteristics. For instance, to consider geographic risks we could add another factor to (3.2) to model p_{ijk} and change c_i to c_{ik} or change $f(s_{ij})$ and $g(l_{ij})$ to $f_k(s_{ij})$ and $g_k(l_{ij})$, where k indicate different regions. As a result, many nonfinancial factors, which are difficult for structural model to consider, can be handled in a similar way by using penalized splines method.

Chapter 4

Investigating Burnout Effect Using Penalized Splines

4.1 Motivation

The heterogeneity in prepayment behavior within a pool has been identified in previous research on prepayment models. It is summarized as the burnout effect in many pieces of literature such as Stanton (1995). In spite of significant academic and practical development in the prepayment modelling, there has been little research directly focusing on the stability of the impact of the burnout effect. In other words, whether or not the impact of the heterogeneity within a pool on the prepayment behavior and valuation remains unchanged over time is not clear. Moreover, how to incorporate this varying impact, if it exists, into prepayment models might be meaningful. In this chapter we explore the stability of the impact of burnout on the prices of mortgage backed securities using penalized splines as estimation method. Before we go to the empirical part we first look at the classical way of considering burnout effect.

Corresponding to a reduced form approach and structural approach respectively in prepayment modelling, the burnout effect can be considered in two different ways. In the structural approach, burnout is explained as the result of heterogeneity in the pool. Hayre (1994) proposes a decomposition of a mortgage pool into slow and fast subpools. These subpools prepay at different speeds, which lead to the burnout phenomenon. Stanton (1995) extends the rational prepayment models

in Timmis (1985), Dunn and Spatt (1986), Johnston and Drunen (1988) by adding heterogeneous transaction costs and prepayment decision at discrete intervals to endogenously produce the burnout effect and finds that different classes of mortgage borrowers prepay at different times due to heterogeneous costs. Levin (2001) applies a two components model, namely, active and passive, to burnout analysis, mortgage valuation and average life analysis respectively.

In the reduced form approach, the burnout effect is usually defined as a variable, which has an impact on the prepayment probability through some hazard function. Survival models or logit models are used in frequently found literatures such Clapp et al. (2006). In terms of burnout effect, a significant difference among these models lies in the way in which burnout effect is measured. For example, Richard and Roll (1989) in defining burnout use an exponent function of the ratio of the mortgage coupon rate to the refinance rate, whereas Schwartz and Torous (1989) use the log value of the proportion between the amount of the pool outstanding and the principle in the absence of prepayment. Schwartz and Torous (1993) model burnout using the sum of the maximums of the difference between mortgage coupon rate and refinance cost and risk free rate. Moreover, they explicitly outlines the relationship between burnout and prepayment probability, which is approximately decreasing. Charlier and Bussel (2003) uses the difference between the refinance incentive in the current month and the maximum refinance incentive as a measure of burnout for the prepayment in the Dutch market.

To directly investigate empirical data on the burnout effect, we here focus on the reduced form prepayment model. As far as burnout is concerned, most of the above reduced form prepayment models share the following two features. On the one hand, they use constant coefficients for the explanatory variables in hazard models. The sign of the coefficient before burnout is typically predicted to be negative according to economic intuition and then verified by the estimation results later like in Schwartz and Torous (1993), Mattey and Wallace (2001), Charlier and Bussel (2003). On the other hand, the burnout effect is typically modelled independently. These aspects of prepayment modelling can be improved upon. A growing literature recognizes the importance of extending the assumption of constant coefficients. Kau and Springer (1992) extend this popular assumption of fixed coefficients in prepayment modelling and use two different random coefficient models

(RCM), namely Swamy RCM[4.1] and Hildreth and Houck RCM[4.2], to model pooled data and individual securities respectively. They verify the randomness of the coefficients and succeed in challenging the fixed coefficient assumption. Kau and Springer (1993) investigate the impacts of financial and nonfinancial incentives on prepayment, based on the assumption of the varying coefficients having polynomial structures and the corresponding estimation results. LaCour-Little and Green (2002) also question the constant coefficients assumption based on technological improvement and greater efficiency in the mortgage market. They verify the instability of parameters by comparing two hazard models estimated for different periods and calculate this impact on pricing mortgage backed securities. Although the above analyses identify the limitation of the assumption of constant coefficients in prepayment modelling, the stability of coefficients of burnout effect is beyond their investigation. Research explicitly highlighting the stability of coefficients includes Popova et al. (2006), who propose a Bayesian mixture model for prepayment rates of individual pools of mortgage. They find that the coefficients of burnout effect can be negative or positive for different pools in their estimated results. In other words, the burnout effect can have negative or positive impact on prepayment rates, which is not in line with the economic intuition. They confirm this estimation result by checking the data but provide no definite conclusions or directions on how to improve prepayment modelling as they focus on the estimation procedure. In addition to the coefficients problem, whether the burnout effect is interacting with other variables is also not clear. This chapter contributes to literature in two perspectives. First, we link the burnout effect directly with the prices of mortgage-backed securities. We take advantage of the capability of the nonparametric approach to capture the nonlinear relationship between prices and burnout. Secondly, we emphasize the interaction between burnout and the life of the MBS, which is represented by scheduled factors. From the interaction we could identify how the burnout effect is affected by scheduled factor.

[4.1] See Swamy (1970)
[4.2] See Hildreth and Houck (1968)

4.2 Measuring Burnout Effect

4.2.1 Burnout

Intuitively, mortgage borrowers who are sensitive to the changes in interest rates will prepay earlier. Thus, the remaining part of the pool consists of the borrowers who are relatively insensitive to the opportunity to prepay. Burnout effect is used to explain this phenomenon. Namely, prepayment is decreasing over time and does not appear even when refinancing incentives are available. To identify the effect of burnout we first need a function to quantify the burnout effect. This function depends on two different factors; observed and scheduled factors. The observed factor (OF) is the ratio between the observed remaining principal outstanding and the original principal outstanding. The scheduled factor (SF) is the ratio between the scheduled remaining principal outstanding and the original principal outstanding. Since the value of the factor depends on the remaining principal it also reflects any changes in the principal induced by the prepayment behavior. Therefore, we choose the following function of factors to measure burnout effect in line with Jegadrrsh and Ju (2000),

$$BO_t = 1 - \frac{OF_t}{SF_t} \tag{4.1}$$

where BO_t measures burnout at time t, OF_t is the observed factor at time t and SF_t is the scheduled factor at time t. A smaller BO means that the observed factor is closer to scheduled factor. In other words, there was less prepayment behavior in the past, which also indicates that prepayment is more likely to happen in the current period and in the future. Conversely, under the same conditions, the larger the burnout of a pool is, the less prepayment behavior may appear in current period.

4.2.2 The Behavior of Factor and Burnout

The series OF and SF decrease over time where the speed of decreasing varies among pools with different coupon rates. Figure 4.2.2 shows an example of scheduled factor series of pools with different coupon rates. Since the mortgage with a higher loan rate will pay more interest, the proportion of principle payment in the total payment will be smaller in case of amortization. Thus, the factor series

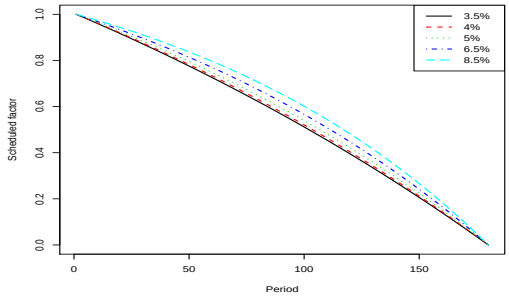
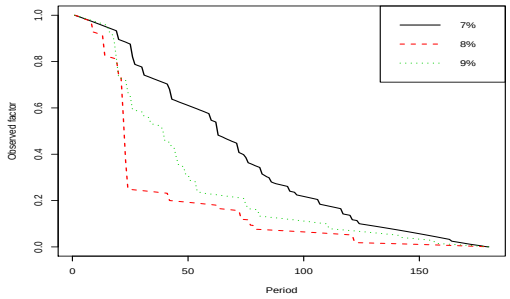


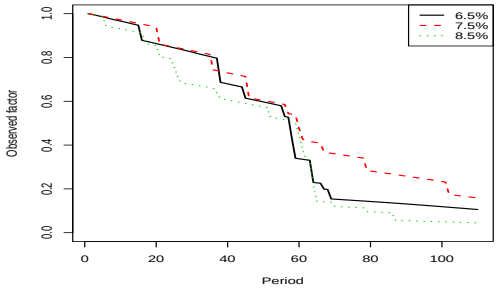
Figure 4.1: Scheduled factors of pools with different coupon rates over ages

measuring remaining principle balance of higher coupon rates will be larger.

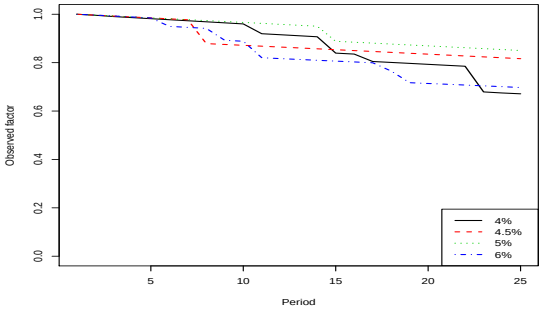
The behavior of the observed factor becomes more complicated as prepayment is considered. Figure 4.3 shows some observed factor series (OF) of pools with different coupon rates and three different issue dates. Pools with such high coupon rates do not have higher factors as in the scheduled case because higher coupon rates may lead to more prepayment behavior. Plot (b) in Figure 2 shows that the factor series of a pool with 8.5% coupon rate experiences the largest drop. In contrast to plot (b), plot (a) shows that a pool with 8% coupon rate has a larger decrease than a pool with 9% coupon rate. The comparison between the two plots results in the conclusion that the coupon rate is only one key reason for the prepayment behavior. As the structural approach demonstrates, a phenomenon named burnout is produced when prepayment decision is affected by heterogenous costs. Therefore, adding burnout effect into our models will better explain the prepayment behavior and the prices. Once we have the scheduled and observed factor series we could calculate the burnout according to (2). Figure 3 illustrates three different burnout series. Generally, the burnout remains small at the early phase, then experiences a climbing process, which results in a relatively stable end phase.



(a) Observed factor series of three pools issued in 1992

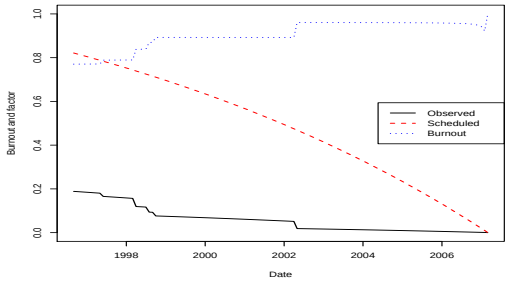


(b) Observed factor series of three pools issued in 1998

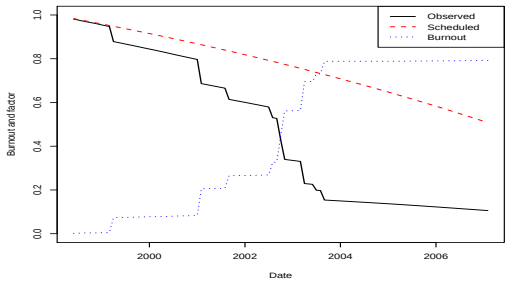


(c) Observed factor series of three pools issued in 2005

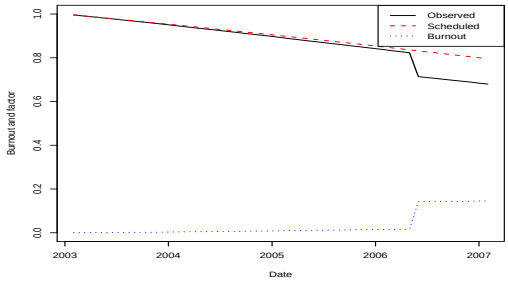
Figure 4.2: observed factors of different pools with different coupon rates issued in different year.



(a) Observed factor series, scheduled factor series and burnout of a pool issued in 1992 with coupon rate 8% in end phase



(b) Observed factor series, scheduled factor series and burnout of a pool issued in 1998 with coupon rate 6% in middle phase



(c) Observed factor series, scheduled factor series and burnout of a pool issued in 2003 with coupon rate 4.5% in begin phase

Figure 4.3: Observed factor series, scheduled factor series and burnout.

4.3 Nonparametric Prices Model

4.3.1 Modelling Prices of MBS with Burnout Effect

As discussed in previous chapter the nonparametric approach has been widely used to investigate MBS prices and prepayment rates. We follow their nonparametric trend, but add more variables to estimate MBS pool prices by using an additive model, which is estimated by penalized splines smoothing technique. We demonstrate that a simple nonparametric model with interaction between burnout and schedule factors is not only more powerful than structural approach in illustrating the varying burnout effect, but is also tractable for specific mortgage pool valuation. We subsequently restrict our attention to variables that are expected to have the most significant impacts on prices. We specify explicitly three components influencing prices of mortgage backed securities; a straight bond, an option component and an interaction between burnout and scheduled factor. The first three parts of the price are mainly related to coupon rates, the short term rate and long term interest rates, as proxy of a straight bond. Since the refinance incentive is related to long term interest rate and hence reflects the prepayment behavior, the long term interest rate plays the main role in the option part. This is in line with Kau, Keenan, Muller, and Epperson (1992), who show that the incentive dominate the prepayment rates. The final part considers the impact of burnout. Here we use the interaction between burnout effect and scheduled factor to incorporate a varying burnout effect. Thus, prices of MBS are modelled as follows,

$$p_{it} = f_1(c_i) + f_2(s_t) + f_3(l_t) + f_4(BO_{it}, schedule_{it}) + \epsilon_{it}, \quad 1 \leq t \leq N_i \quad (4.2)$$

where p_{it} is the price of the i-th MBS at time t, c_i is the coupon rate of i-th MBS, l_t and s_t are long term and short term interest rates, BO_{it} measures burnout at time t, $schedule_{it}$ is the scheduled factor of i-th at time t, f_1, f_2 and f_3 are three univariate unknown smoothing functions, f_4 is a bivariate unknown smoothing function, ϵ_{it}, $t = 1, \ldots, N_i$ are drawn from a $AR(1)$ process, which is assumed to be correlated within one MBS. The intuition behind (4.2) is as follows. The role of coupon rate, long term and short term interest rates in the prices of MBS are captured by unknown function f_1, f_2 and f_3. To identify the impact of maturity on the burnout,

we set a bivariate function f_4. There are two benefits of adding f_4. First, the scheduled factor and burnout are modelled in an interactive way. This makes it possible to explicitly model the impact of burnout on prices. Usually, the burnout effect is assumed to have a negative impact on the prepayment behavior, and in turn a positive impact on prices. Spahr and Sundermann (1992) shows that the burnout effect has positive impact on the prices of MBS by comparing Goldman Sachs prepayment models[4.3], with or without burnout effect being used respectively. To test and extend their conclusion, the burnout effect is now restricted by scheduled factor. This gives more flexibility to model the price response to burnout changes. This is because, the burnout does not have to show the same impacts on prices in different maturities. The second benefit of adding f_4 is that, this model can then be easily extended to model possible unobserved heterogeneity between pools with different maturities. The main part of heterogeneity comes from geographic diversity, differences in education level, coupon rates, issue date and other observable characteristics. Stanton (1996) shows that heterogeneity can be extracted from the information about prepayment behavior over time. In our model, the heterogeneity between pools results in each pool having different values for the burnout effect, BO_{it}, over time. Continuing to assume that prices for each pool is determined by different unknown functions in (4.2), we can model heterogeneity within a pool or between pools.

4.3.2 Estimation Methodology

Nonparametric technique is used to uncover the relationships between these financial variables. Here we estimate these unknown functions by penalized splines, which is discussed in Chapter 2. An advantage of penalized splines is its link to linear mixed model. This feature allows to use the software developed for mixed model can be used to estimate the parameters in a penalized splines model. Moreover, the link to mixed model makes it easier to cope with smoothing parameter selection in case of correlated errors, which is shown in Krivobokova and Kauermann (2007). We will take advantage of the latter subsequently. Numerically, the fit can be calculated with available software using R. In particular, all fits have been calculated with standard setting using the procedure **gamm()** provided in package **mgcv**, see

[4.3] See Richard and Roll (1989)

Figure 4.4: 10-year treasury note yield and 3-month treasury bill yield

also Wood (2006). More details are provided in Appendix B and Appendix D.

4.3.3 Empirical analysis

Data

We now estimate the four parts in Model (4.2) using data from different pools. Our data are collected from Reuters 3000 Xtra. It consists of the pool-specific information such as prices and factors of 68 single family pools over the period from August 1996 to February 2007. The pools discussed in this study are backed by guarantees provided by the Government National Mortgage Association. Due to this insurance feature, default appears as prepayment behavior. However, in contrast to other structural approaches concerning house prices like Mattey and Wallace (1998) and Downing et al. (2005), we will not include house prices in our analysis for two reasons. Firstly, the house price index increases over our observation period and the default behavior is more evident in times of decreasing house prices. Secondly, we do not have a specific house price index related to each pool.

Table 4.2 provides the structure of our data. It shows a decreasing trend in issue coupon rates during the period from 1992 to 2006. This trend also corresponds to the dynamics in risk free interest rates such as the 10-year yield of Treasury note and 3-month yield of Treasury bill, which is shown in Figure 4.4. Figure 4.5 shows the observed prices of 68 pools. Consistent with the economic intuition, the prices illustrate opposite movement to the interest rates.

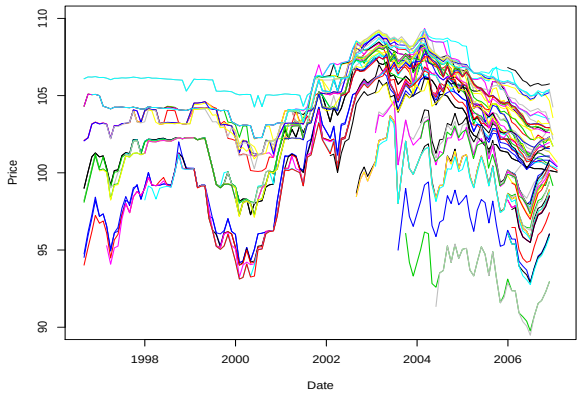


Figure 4.5: Prices of 68 pools over the period from August 1996 to February 2007

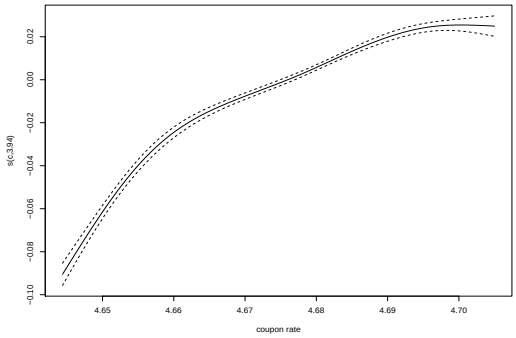


Figure 4.6: Estimated coupon rate component $f_1(c_t)$

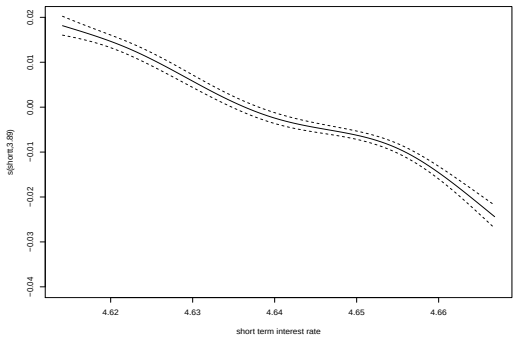


Figure 4.7: Estimated short term interest rate component $f_2(s_t)$

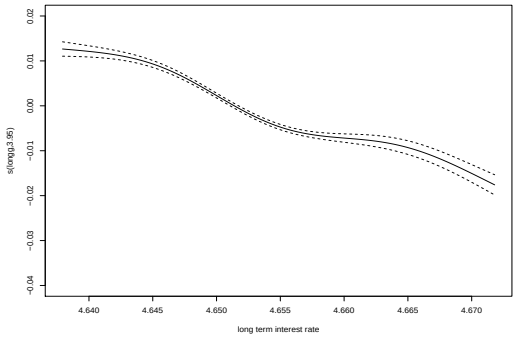


Figure 4.8: Estimated long term interest rate component $f_3(l_t)$

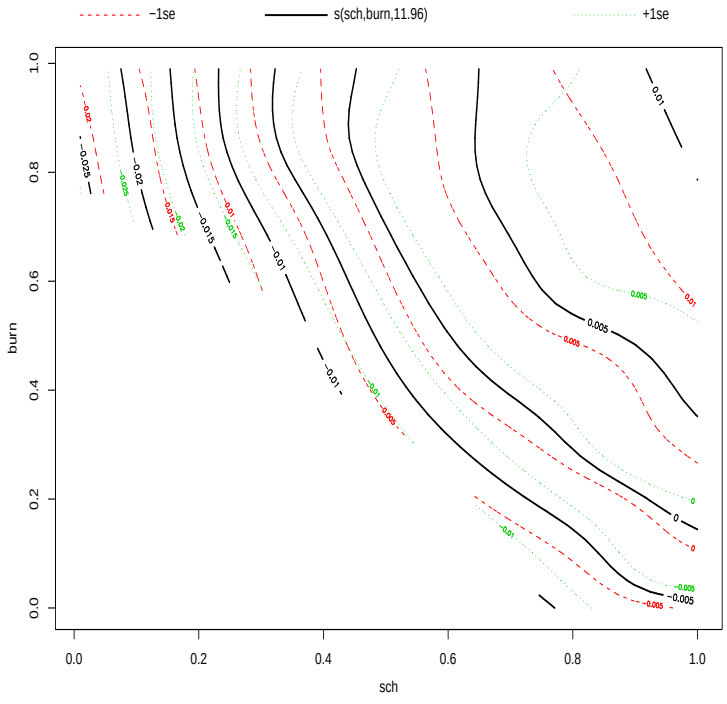


Figure 4.9: The interaction between scheduled factors and burnout

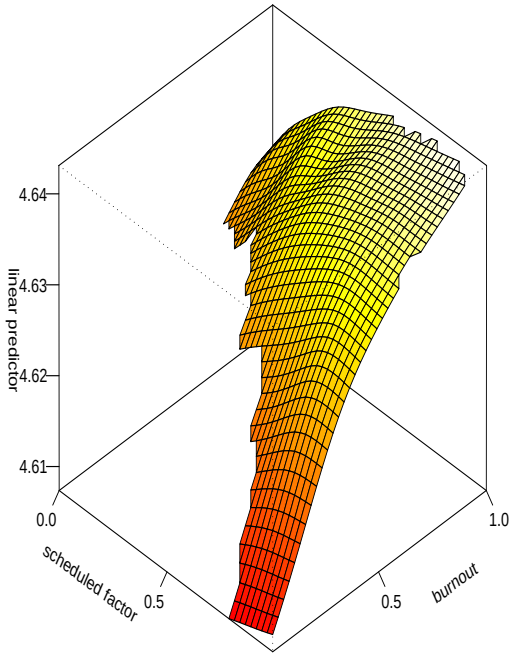


Figure 4.10: 3D illustration of the interaction between scheduled factors and burnout

	Estimate	Std Error	t value	Pr($>\|t\|$)
Intercept	4.6345179	0.0007722	6001	<2e-16

Approximate significance of smooth terms:

	edf	Est.rank	F	p-value
s(c)	3.938	4	462.181	$<$ 2e-16
s(longg)	3.952	4	166.133	$<$ 2e-16
s(shortt)	3.888	4	143.201	$<$ 2e-16
s(sch,burn)	11.960	24	3.818	<1.01e-09
R-sq.(adj) =	0.865	n =	4456	

Table 4.1: Estimation result

Result

The nonparametric model (4.2) also captures the main part of the price changes. Table 4.1 shows the extracted information about the results of the estimation. 86.5% of the variance is explained by model (4.2). We can see that the first three components of fit of (4.2) are consistent with the economic intuition: the prices are decreasing with long term and short term interest rates and increasing with coupon rates.

Figure 4.9 and Figure 4.10 show the estimated interaction of the scheduled factors and burnout in two dimension and three dimension, respectively. The estimated interaction between scheduled factors and burnout exhibits two important features. First, the relationship between prices and burnout is varying with scheduled factors. The higher the scheduled factor the stronger the effect of burnout. This means that our assumption of the interaction between burnout and refinance incentive is reasonable. Secondly, for most part of the surface, the estimated relationship between burnout effect and prices is increasing. The property of increasing is more obvious in areas with larger scheduled factor, in other words, at the beginning phase of the pools. This is in agreement with the economic meaning of burnout. It measures the contribution of refinance incentive to the prepayment behavior. At the beginning phase, there is a rapid increase in prepayment rate, therefore the burnout effect linked to refinance incentive should also be significant. After the middle stage burnout has little impact on the prices and the surface becomes more flat. To confirm the observation that prices are not decreasing with burnout we modify the price

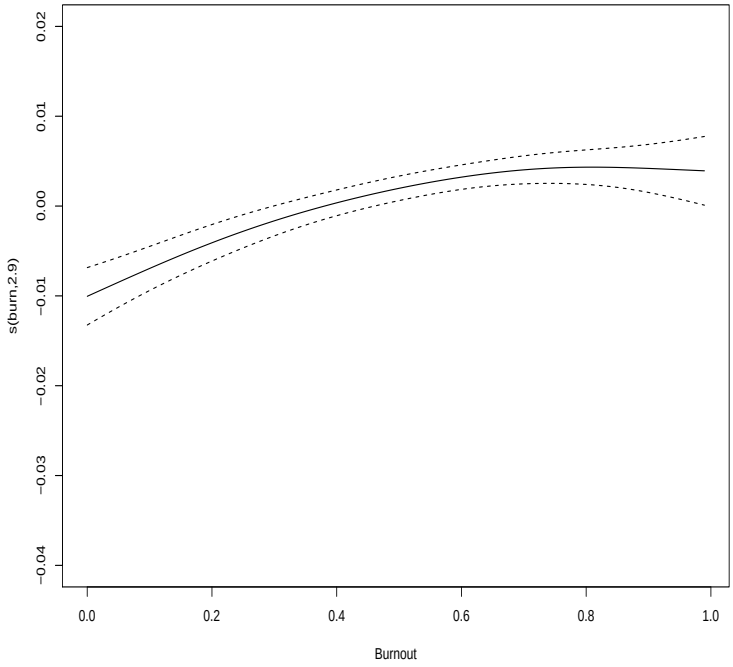


Figure 4.11: Estimated univariate function of burnout

model by decomposing the bivariate part into two univariate functions. Figure 4.11 shows the estimated univariate function of burnout.

4.4 Conclusion and future extensions

In this chapter we focus on investigating the impact of burnout effect on the prices of mortgage-backed securities. We first model the prices of mortgage-backed securities by the linear combination of several unknown functions and then estimate them using penalized splines method. The results show that the impact of burnout effect on the MBS prices can be divided into two stages, the evident part in the earlier stage and the flat part in the later stage. If sufficient prepayment data is available, future research should include the empirical part of prepayment modelling with burnout effect using penalized splines.

Maturity			Coupon	Rate					
2007						7	8	9	
2008					6	7	8	9	
2009				5	6	7	8	9	10
2010					6	7	8	9	10
2011					6	7	8	9	
2012					6	7	8	9	
2013		4	4.5	5	6	7	8		
2014			4.5	5	6	7	8		
2015				5	6	7	8		
2016			4.5	5	6	7	8		
2017		4	4.5	5	6	7			
2018	3.5	4	4.5	5	6				
2019	3.5	4	4.5	5	6				
2020		4	4.5	5	6				
2021			4.5	5	6				

Table 4.2: The structure of pool data

Chapter 5

Exploring the Credit Risk for Small and Medium Enterprises in China

5.1 Introduction to Credit Risk

Credit risk is the risk of loss as a result of unexpected changes in the credit quality of a borrower or counterparty in a financial agreement. Credit risk models can be classified into two categories, one is used to price a single credit sensitive security, and the other is used to investigate the credit risk of a portfolio. In both situations, credit risk is analyzed by either structural models or reduced form models. Structural models are based on the work of Black and Scholes (1973) and Merton (1974) and their extension, Hull and White (1995). In these models, default is a result of the feature variables of a firm, such as asset value, asset volatility and capital structure. Industry examples of credit portfolios include J. P. Morgan's CreditMetrics^{TM} and KMV'S CreditPortfolioManager^{TM}. Reduced-form models date back to Jarrow and Turnbull (1995) and Duffie and Singleton (1999). The default is modelled as a stochastic process linked to some exogenous random variables like macroeconomic variables, rather than the characteristic variables included in structural models. Pricing is implemented by a default-risk-adjusted interest rate with intensity considered. Industry examples for credit portfolios include Credit Suisse First Boston's CreditRisk^{+} and McKinsey's CreditPortfolioView. Further details on the credit risk modelling are included in Bluhm et al. (2003), Bielecki and Rutkowski (2002), or Schönbucher (2003).

Credit risk models are becoming popular due to the implement of Basel II accord. Basel II is a comprehensive framework for regulatory capital and risk management. Under this framework, a bank is required to maintain capital equal to not less than 8% of the aggregate value of its risk-weighted assets (RWA). The capital charge on a portfolio can be obtained by the sum of the capital charges on a loan-by-loan basis.

$$RC = 0.08RWA = \sum_{i=1}^{n} 0.08RWA_i = \sum_{i=1}^{n} RC_i$$

where RC is the aggregate risk capital, RC_i and RWA_i are the risk capital and the risk weighted asset of obligator i respectively. In IRB approach, the risk capital is calculated as

$$RC_i = c\ \delta_i\ E_i\ \phi(\frac{\phi^{-1}(p_i) + \sqrt{\rho_i}\phi^{-1}(\alpha)}{\sqrt{1-\rho_i}}) \tag{5.1}$$

where c is an adjustment parameter, δ_i is the percentage loss given default, E_i is the exposure, ρ_i is a correlation parameter measuring dependency, p_i is the marginal default probability, ϕ is the cumulative distribution function of the standard Gaussian distribution and α is the confidence level, which is set to 99.9% in Basel II.

As shown in (5.1), δ_i, E_i, ρ_i and p_i are parameters to be determined for each obligator, of which p_i is more central in credit risk modelling than others. The default probabilities can be obtained either from an external source like the publication of rating agencies such as S&P and Moody's or internal assessment, in which case statistical models, such as logit or probit-models, are used to assign default probabilities to non-rated obligors.

5.2 Motivation of Credit Risk Modelling for SMEs in China

In the Chinese credit markets there is a gap between small and medium size enterprises (SMEs) and banks. Each side faces its own dilemma. SMEs become increasingly important to the rapidly growing Chinese economy. In October 2006, the number of SMEs in China was over 40 million, 99.6% of all enterprises. SMEs

also account for 58.5% of the gross domestic product and 75% of urban employment (See Shanghai Securities News (2006)). However, most of them are suffering from the lack of access to finance. Hussain et al. (2006) and reports from the World Bank and the Peoples's bank of China show that Chinese SMEs depend largely on the internal source for financing instead of external sources, furthermore, the external bank credits are mainly short term. The main reasons for Chinese SMEs' difficulties in accessing the bank credit can be classified into two categories. First, there are some typical characters of SMEs, such as, less external ratings, less financial and operational transparency, obscure financial statements, the family-owned nature of many SMEs and a short history of the relationship with banks. Beck and Demirguc-Kunt (2006) also show that financial institutions play an important role in relaxing the growth constraint of SMEs due to lack of access to finance. The other reason is mainly due to the transition process of Chinese economy from 1978, which is featured by an increasing market competition, changing policies and the integration into global economy.

Meanwhile, Chinese banking industry has been open to foreign banks since January 2007. Clarke et al. (2005) shows that Latin America foreign banks with large local presence lend more to small business. In the past decades relationship and lending has been based on the information gathered through business over time and a case by case decision. However, this lending technique is expensive and time consuming. Thus, Chinese banks face the dilemma, whether they should ignore the growth of SMEs and implement a profit-oriented strategy requiring them to focus more on large cities and large-scale enterprises, or they should be more active in SMEs markets irrespective of cost consideration as a response to the competition between domestic and foreign banks. To solve this dilemma, Chinese banks could turn to alternative quantitative lending techniques for SMEs. Unlike the listed companies, the market value of SMEs is usually unavailable. Hence, the credit models based on the ideas from firm's value theory introduced by Black and Scholes (1973) and Merton (1974) can not be applied to SMEs. Thus, the methods used to model consumer credit, such as credit scoring, are more appropriate to measure the credit quality of SMEs.

Credit scoring is a statistical method used to evaluate the credit risk. It is important for Chinese banks for the following reasons. First, Blöchlinger and Leippold (2006) show that a good credit scoring model can increase banks' ability to identify

risk, implement suitable pricing strategy and hence increase profits. Second, Berger and Frame (2007) show that credit scoring can also contribute to the availability of credit for SMEs. Finally, credit scoring model also enable banks to have less capital in addition to the above advantages. The Basel 2 Accord explicitly requires banks to calculate the probability of default in case the bank determines the capital using internal ratings based approaches instead of the standard approach. An advantage of the former is that it results in less capital requirement for banks.

Credit scoring models are widely accepted in consumer credit industry. The field of credit scoring is thereby dynamic and new statistical techniques are being used. Statistical methods used in credit scoring usually include logistic regression, probit regression, discriminant analysis, neural network, genetic algorithms, linear programming, K-nearest-neighbor classifier and classification trees. Hand and Henley (1997) provide a summary of these methods in consumer credit scoring. Tomas et al. (2005) provide further discussion. Leonard (1992) shows that the implementation of such credit scoring models for small business loans is feasible as well. Among these consumer credit scoring methods the logistic model is quite popular because of its good balance between simplicity and accuracy. Altman and Sabato (2007), Behr and Guttler (2007), Phillips and Vanderhoff (2004) show its application to predict loan default. The impact and recent developments in credit scoring is discussed in Hand (2005) and Crook et al. (2007).

A successful credit scoring model can only work if the data behave somewhat stationary, that is to say that credit behavior today gives information about credit behavior in the future. From a statistical point of view this means that covariate effects which influence the probability of loan default should not change with (calender) time. Apparently in times of economics transitions this assumption is likely to be violated and should be replaced by models where covariate effects vary with time. Models of this kind are known in statistics under the phrase varying coefficient models, see Hastie and Tibshirani (1993), Wood (2000) or Ruppert et al. (2003). This model class has become quite popular in the last decade. We make use of this tool and investigate if and how covariate effects have changed in the last years, that is how the economic transition is mirrored in the credit risk. To do so, we extend Müller and Rönz (2000) and Liu and Cela (2006) in three aspects. First, we use several univariate nonparametric functions to model the risk of default in a nonparametric style. Secondly, we apply time varying coefficients in linear parts,

that is we show explicitly how the dynamics of the Chinese banking industry is mirrored in varying effects of risk factors. This change accounts for the transition of the Chinese economy over the last ten years. Finally, for fitting we make use of penalized splines, which overcomes the overfitting problem usually occuring with the high dimensional techniques and provides a simple model selection idea.

The data set for this analysis includes 1379 commerical loans from a Chinese bank over the years 1998-2005. The data are available from the authors, the name of the bank however has to be remain confidential according to the agreement with the bank. The variables consist of risk status, entry time, enterprise types, guaranty methods and loan amount.

5.3 Credit Scoring with varying coefficients

5.3.1 Credit Scoring Model

Assume that $y \in \{0, 1\}$ is an indicator for default. We model the probability of the default as a function of selected independent variables. This allows us to classify the loan to be good ($y = 0$) or bad ($y = 1$) in credit scoring. A simple approach for doing so is the logit model,

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}} \tag{5.2}$$

where $X = (x_1, \cdots, x_k)$ is a vector of covariates. The model is widely used in practice and classification of whether an existent or new credit is reliable depends on whether the probability in (5.2) exceeds a given threshold.

Model (5.2) by itself is parametric and therefore static and does not mirror dynamic effects. This is what we pursue now by investigating how the dynamics in the Chinese economy influences the probability of default. Due to the development of nonparametric methods in the last two decades the linear part in (5.2) is extended to be nonlinear. Notice that the parameters in (5.2) are independent of entry time. Therefore, no matter when a loan is issued, the predictors have the same effect on the credit score. This is questionable. If a company borrows in a different entry time, the credit score should also change correspondingly. This change is particularly necessary for SMEs in a transition economy like China because the

macroeconomic environment is changing dramatically. In the last decades China has been experiencing the transition from a central planning to a market economy. Thus, the entry time should play an important role for the performance of the SMEs loans. Similarly, including time-dependent coefficients in the credit scoring can not only reflect the evolution of the effect of different predictors but also consider the stability of parameters over time. We realize the consideration of the entry time by using a varying-coefficient model.

Let $\mathcal{J}$ be the index set of categorical variables considered including the intercept, and let $\mathcal{I} \subset \mathcal{J}$ be the subset of these variables having time varying effects including the intercept. Moreover, we denote with $\mathcal{C}$ the index set of continuous, metrically scaled variables and $\mathcal{D} \subset \mathcal{C}$ are the indices of continuous variables interacting with time. A generalized logit model with time varying effects for credit scoring is then as follows,

$$P(Y = 1 \mid X, t) = F\{\sum_{j \in \mathcal{J} \setminus \mathcal{I}} \beta_j x_j + \sum_{j \in \mathcal{I}} \beta_j(t) x_j + \sum_{j \in \mathcal{C} \setminus \mathcal{D}} f_j(x_j) + \sum_{j \in \mathcal{D}} f_j(x_j, t)\} \tag{5.3}$$

where $F(\eta) = \frac{1}{1+exp(-\eta)}$ is the logistic distribution function and t represents the calender time. In (5.3) $f_j(x_j, t)$ are bivariate functions while $f_j(x_j)$ depend on x_j only. Analogously, $\beta_j(t)$ gives the effects of covaraiate x_j which are dynamically changing with time while β_j is a constant effect of covaraiate x_j. It is easy to see that model (5.3) is not identifiable. The usual assumption is therefore to postulate that the functional components in (2) integrate out to zero, e.g. $\int f_j(x_j, t)dx = 0$ or $\int f_j(x_j, t)dt = 0$. Finally, model (5.3) has a sophisticated structure, thought it might not be parsimonious. Therefore, model selection will be necessary which will be pursued later on.

5.3.2 Estimation Methodology

Penalized splines have become a popular smoothing technique over the last couple of years, originally introduced by O'Sullivan (1986), it were Eilers and Marx (1996) who demonstrated the simplicity and efficiency. Its link to the mixed model framework has been demonstrated in Ruppert et al. (2003) and Wood (2006). We take advantage of these developments. Let us first briefly introduce the penalized

spline method. We replace the smooth functions $f_j(x_j)$ and $\beta_j(t)$, respectively, by high dimensional basis functions in the form

$$f_j(x_j) = \tilde{B}_j(x_j)\tilde{u}_j, \qquad \beta_j(t) = B_j(t)u_j$$

where $\tilde{B}_j(x_j)$ and $B_j(t)$ are high dimensional basis functions built e.g. from vectors of B-splines or truncated polynomials (see Ruppert et al. (2003) and Wood (2006)) and $\tilde{u}_j$ and u_j are the corresponding parameters. By analogy with the univariate case we replace the two dimensional function with a high dimensional basis

$$f_j(x_j, t) = \bar{B}_j(x_j, t)\bar{u}_j$$

where $\bar{B}_j(x_j, t)$ is a basis in two directions. This gives a high dimensional parametric model with parameters, β_j with $j \in \mathcal{J}\backslash\mathcal{I}$. u_j with $j \in \mathcal{I}$, $\tilde{u}_j$ with $j \in \mathcal{C}\backslash\mathcal{D}$ and $\bar{u}_j$ with $j \in \mathcal{D}$. Apparently, fitting will be unstable due to the high dimensionality, so that we include penalty terms on the coefficients in the form, $\lambda_j u_j^T D_j u_j$ with $j \in \mathcal{I}$, $\tilde{\lambda}_j \tilde{u}_j \tilde{D}_j \tilde{u}_j$ with $j \in \mathcal{C}\backslash\mathcal{D}$ and $\bar{\lambda}_j \bar{u}_j^T \bar{D}_j \tilde{u}_j$ with $j \in \mathcal{D}$. Where $\lambda_j, \tilde{\lambda}_j and \bar{\lambda}_j$ steer the amount of penalization. Since this has a direct effect on the smoothness of the fit we also call them *smoothing parameters*.

Let $l(\beta, u)$ be the likelihood based on the data, with $\beta = (\beta_j, j \in \mathcal{J}\backslash\mathcal{I})$ and u as vector containing u_j, $\tilde{u}_j$ and $\bar{u}_j$ stacked up to a vector. Then the penalized likelihood takes the form,

$$l(\beta, u, \lambda) = l(\beta, u) - \frac{1}{2}\sum_{j\in\mathcal{I}} \lambda_j u_j^T D_j u_j - \frac{1}{2}\sum_{j\in\mathcal{C}\backslash\mathcal{D}} \tilde{\lambda}_j \tilde{u}_j^T \tilde{D}_j \tilde{u}_j - \frac{1}{2}\sum_{j\in\mathcal{D}} \bar{\lambda}_j \bar{u}_j^T \bar{D}_j \bar{u}_j \tag{5.4}$$

Estimation of the smoothing parameters can be done by treating the penalties as normal priors, e.g. $u_j \sim N(0, \sigma_j^2 D^*)$, $\tilde{u}_j \sim N(0, \tilde{\sigma}_j^2 \tilde{D}^*)$, $\bar{u}_j \sim N(0, \bar{\sigma}_j^2 \bar{D}^*)$, D^*, $\tilde{D}^*$ and $\bar{D}^*$ as generalized inverse of D, $\tilde{D}$ and $\bar{D}$, respectively, and $\sigma_j^2 = 1/\lambda_j$, $\tilde{\sigma}_j^2 = 1/\tilde{\lambda}_j$ and $\bar{\sigma}_j^2 = 1/\bar{\lambda}_j$. The model now becomes a Generalized Linear Mixed Model (GLMM) and the smoothing parameters can be derived as maximum likelihood estimates. The procedure is implemented in R following the package SemiPar or the newest version of the $gamm()$ routine in package mgcv for non-normal responses.

5.3.3 Model selection by using the marginal likelihood

Model selection relates to the exercise of selecting the four set $\mathcal{D}$, $\mathcal{C}$, $\mathcal{I}$ and $\mathcal{J}$ in model (2) based on the data. In particular, we need to determine which of the covariate interact with time, which determines how subsets $\mathcal{I} \in \mathcal{J}$ and $\mathcal{D} \in \mathcal{C}$ are chosen appropriately. To tackle the exercise, we consider the marginalized likelihood, that is after integrating out the random spline effects. Based on a Laplace approximation this results to

$$l_{\mathcal{M}}(\beta, \lambda) := \log \int exp\left(l\left(\beta, u\right)\right)\phi(\boldsymbol{u}, \boldsymbol{D}(\lambda))d\boldsymbol{u}$$

$$\approx l(\beta, \hat{u}, \lambda) + \frac{1}{2}\log|(\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{B} + \boldsymbol{D}(\lambda))^{-1}\boldsymbol{D}(\lambda)| \tag{5.5}$$

where λ is the vector of coefficients λ_j, $\tilde{\lambda_j}$, $\bar{\lambda_j}$; $\boldsymbol{W}$ is the weight matrix containing the binomial variances and $\boldsymbol{B}$ is the entire basis matrix with i-th row constructed from $((B_j(t_i),\ j \in \mathcal{I}),\ (\tilde{B}_j(t_i),\ j \in \mathcal{C}\backslash\mathcal{D}),\ (\bar{B}_j(x_{ij}),\ j \in \mathcal{D}))$ where x_{ij} is the i-th observation of covariate j, $i = 1, \ldots, n$. Finally, $\boldsymbol{D}(\lambda)$ is the block diagonal matrix built from $\boldsymbol{D}(\lambda) = diag((\lambda_j D_j,\ j \in \mathcal{I}),\ (\lambda_j \tilde{D}_j,\ j \in \mathcal{C}\backslash\mathcal{D}),\ (\lambda_j \bar{D}_j,\ j \in \mathcal{D}))$. Coefficient $\hat{u}$ in (5.5) is the maximizer of $l(\beta, u, \lambda)$ keeping β and λ fixed. Maximizing $l(\beta, \lambda)$ now with respect to β and λ yields the penalized fit of the cureves in the model. As sketched in the Appendix it can be shown that (5.5) can be approximated with

$$l_{\mathcal{M}}(\hat{\beta}, \hat{\lambda}) \approx\ l(\hat{\beta}, \hat{u}, \hat{\lambda}) - edf$$

where edf is the estimated degree of freedom calculated from the trace of the smoothing matrix. That is edf gives the estimated complexity of the model with index sets $\mathcal{M} = (\mathcal{C}, \mathcal{D}, \mathcal{I}, \mathcal{J})$ fixed. Let $\hat{\beta_{\mathcal{M}}}$ be the estimate based on index set $\mathcal{M}$ and define $l_{\mathcal{M}}(\hat{\beta}_{\mathcal{M}}) = l_{\mathcal{M}}(\hat{\beta}_{\mathcal{M}}, \hat{\lambda}_{\mathcal{M}})$, where $\hat{\lambda}_{\mathcal{M}}$ is the vector of fitted smoothing parameters based on model with index set $\mathcal{M}$. The intention is now to minimize AIC with respect to the best index set combination. This will be carried out with a forward selection as described in the next section.

State-owned	*Collective-owned*	*Limited-liability*	*Private*	*Others*
121	259	736	118	145

Table 5.1: Enterprise type

5.4 Application to China SMEs data

5.4.1 Data Description

The empirical analysis is based upon small and medium enterprises commercial loan history data from a commercial bank in China over the years 1998-2005. The dependent variable for our analysis is the risk status of the loan measured at the maturity, which has binary values 1 for non-performing and 0 for performing loans. As risk factors we include variables available for all loan contracts. These include loan amount, the entry time, guaranty method, type of enterprise and maturity. Figure 5.1 illustrates the entry time, the duration and the risk status of the data using the lexis diagram. The solid points at the end of the lines indicate that the loans are in non-performing status. Most loans are provided for 1 year. We therefore replace maturity as influencing factor by a discrete covariate indicating short term loans with maturity less than 0.9 and long term loans with maturity larger than 1.1. Figure 5.2 shows the histograms of entry time and loan amounts. For the analyzes we adjust the amount by the yearly fixed-base consumer price index (CPI) from National Bureau of Statistics of China and then take the logarithm as covariate.

The loan applicants are classified into five categories in accordance with the Regulation of the People's Republic of China on the Management of Registration of Corporate Enterprises. Table 5.1 shows the numbers of enterprise types. When signing the contract, many borrowers may provide guarantary methods such as third-party guarantors, mortgages or pledges. Table 5.2 shows the numbers of corresponding guaranty methods. Based on maturity we also define short term and long term loans in addition to the majority of observations. Thus, we have the following explanatory variables concerned in this study, *entry year, log amount, state, collective, private, limited, credit, guarantee, mortgage, pledge, short term* and *long term*.

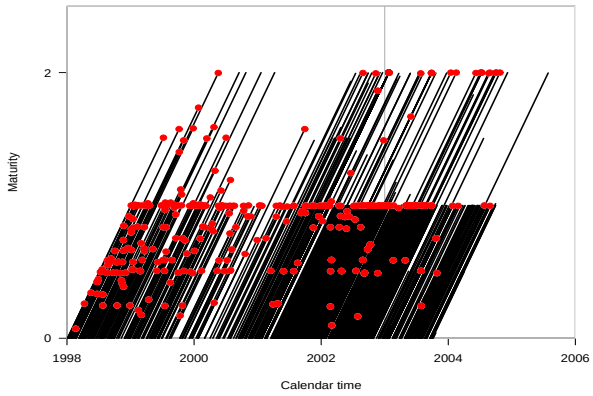
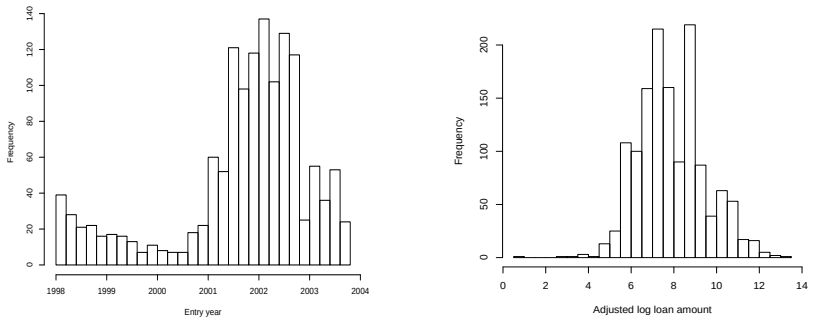


Figure 5.1: Lexis diagram for loans data.



(a) (b)

Figure 5.2: Histograms of entry time and log loan amount.

Credit	*Third-party Guarantor*	*Mortgage-backed*	*Pledge-backed*	*Others*
19	622	664	61	13

Table 5.2: Classification according to guaranty methods

Model	*Potential Combination*	$-2l_{\mathcal{M}}(\hat{\beta}_{\mathcal{M}})$
1	f(entry year) + f(log amount)	979.9257
2	f(entry year, log amount)	977.1435

Table 5.3: Model selection for interaction with time

5.4.2 Model Selection

For model selection procedure we follow a forward-backward strategy as proposed in Edwards and Havránek (1987). The main steps are as follows. First, we look for a semi-parametric model with nonparametric interaction between time and potential continuous covariates, that is we first select index set $\mathcal{D}$. Then, we determine the categorical variables which have time varying effects, that is we find an appropriate index set $\mathcal{I}$. Finally, the categorical covariates with constant coefficients are considered, that is, we look whether we can drop elements of $\mathcal{J}\backslash\mathcal{I}$.

Selection for Interaction with Time

We start with models in a form that an interaction and an unknown function exist for two continuous covariates, here entry year and log loan amount. The other categorical variables are assumed to have constant coefficients at this stage, that is $\mathcal{I} = \phi$. As shown in Table 5.3, we tend to identify the best nonparametric way to incorporate the continuous variables by comparing the $l_{\mathcal{M}}(\hat{\beta}_{\mathcal{M}})$ values of model 1 and model 2. The selection turns out that model 2 including the interaction between entry time and log loan amount has the larger $l_{\mathcal{M}}(\hat{\beta}_{\mathcal{M}})$ value and is therefore chosen.

Selection for Time Varying Coefficients

Now that we have determined the interaction part for continuous variables, we look for the categorial covariates with time varying effects. To do so, we assume that each categorical variable could have time varying coefficients. To decrease the numerical effect of the forward-backward step, we first look for an order of the significances of the time varying effects by repeating the procedure of estimating a model extended from model 2 with the assumption that only one categorial variable has time varying effect. The estimate result is shown in Table 5.4 ordered based on the

values of $l_{\mathcal{M}}(\hat{\beta}_{\mathcal{M}})$. Based on this ordering we now add the time-varying effects of the candidates in Table 5.4 to model 2 one by one until $l_{\mathcal{M}}(\hat{\beta}_{\mathcal{M}})$ does not increase anymore. The process is illustrated in Table 5.5. The result of this model selection shows that we can achieve larger marginal likelihood if we consider time-varying coefficients for short term and long term loan.

Selection for Variable with Constant Coefficients

The final step is to determine the variables with constant coefficients. This is relatively easy. We first estimate a model extended from model 4 (see Table 5.5) with the remaining variables and then choose candidates to be excluded by checking the significance. Each exclusion is also confirmed by the improvement of $l_{\mathcal{M}}(\hat{\beta}_{\mathcal{M}})$. The result of this step shows that our best model consists of an interaction between entry year and log loan amount, short term and long term with time-varying effects and a linear combination of the remaining variables with significant constant coefficients,

$$\eta = \beta_0 + f_1(entry\ year, log\ loan\ amount) + f_2(entry\ year) \times short\ term \\ + f_3(entry\ year) \times long\ term + \beta_1 mortgage + \beta_2 pledge + \beta_3 collective + \\ \beta_4 private + \beta_5 long\ term$$

where η is the predictor in (5.3); Note, $enty\ year$, $log\ loan\ amount$, are metric variables, denoting entry time, log loan amount, respectively, $short\ term$, $long\ term$, $mortgage$, $pledge$, $collective$ and $private$ are dummy variables for short term loan, long term loan, mortgage, pledge, collective-owned enterprises and private companies. The parametric estimates and nonparametric estimates are listed in Table 5.6 and Table 5.7. The graphical illustration of the varying effects is shown in Figure 5.3 and Figure 5.4.

We aim at developing a suitable credit scoring model for the SMEs in China. By doing so the estimation results mimic the path of the Chinese economy and banking industry over the last decade. First, in the model selection procedure, two variables representing the guaranty methods of third-party guarantor and credit are excluded from the model. It indicates that credit and third-party guarantor do not play an important role in determining credit risk. Conversely, mortgage is most statistically significant. As expected, the signs of mortgage and pledge are negative, which means that both guaranty methods decrease credit risk.

Variables	$-2l_{\mathcal{M}}(\hat{\beta}_{\mathcal{M}})$	*Rank*
long term	954.3074	1
short term	968.9114	2
mortgage	976.8645	3
guarantee	977.179	4
state	977.493	5
limited	978.0913	6
collective	978.2383	7
pledge	978.3365	8
credit	978.7852	9
private	978.942	10

Table 5.4: Rank of potential varying effects

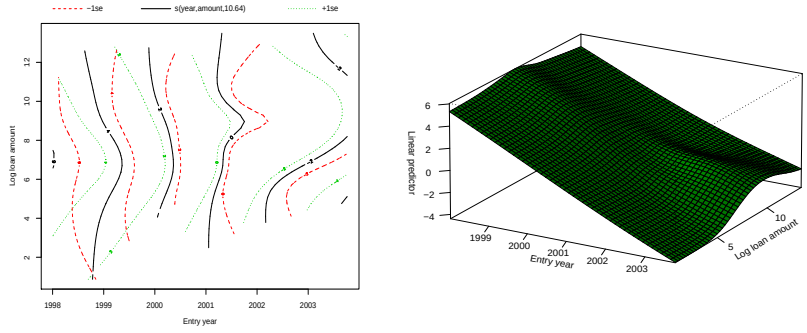
Second, different enterprise types have different effects on the credit risk. Comparing the coefficients in Table 5.6, the loans issued to collective and private companies are more likely to become non-performing loans than state-owned companies. This phenomenon verifies the point of Li and Mehta (2001) that issuing loans to private enterprises does not ensure better performance for banks.

Third, Figure 5.3 shows that most of the credit risk of the same loan amount along the entry time decrease over our observation period from 1998 to 2004. Since the loan amounts have already been adjusted by CPI before estimation, the decrease reflects the improvement of loan performance over years. The shock from the Southeast Asian crisis in 1997 and the membership of World Trade Organization (WTO) in 2001 stimulated Chinese banks to focus on risk management. Figure 5.3 also shows that, for the loans issued the same year, a larger loan amount does not mean the larger credit risk.

Finally, the coefficient of long term in Table 5.6 is the time constant effect while the coefficient of short term is not significantly different from 0 and hence has been excluded. The shapes of the time-varying effects in Figure 5.4 shows that there is no significant time varying impact of short term maturities on credit risk while long term maturities have. The credit risk of a long term loan is always increasing over the last decade.

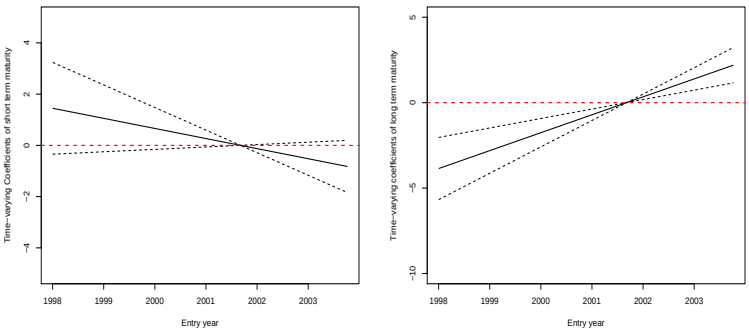
Model	*time varying*	$-2l_{\mathcal{M}}(\hat{\beta}_{\mathcal{M}})$
3	$f_2(t)\times$ long term	954.3074
4	$f_2(t)\times$ long term + $f_3(t)\times$ short term	952.8173
5	$f_2(t)\times$ long term + $f_3(t)\times$ short term+ $f_4(t)\times$ mortgage	954.3189

Table 5.5: Time varying effect selection



(a) Two dimension (b) Three dimension

Figure 5.3: The effect of entry year and log loan amount shown with isolines (left) and three dimensional(right).



(a) Short term (b) Long term

Figure 5.4: Varying effect of short term and long term maturities.

	Estimate	Std. Error	p-value
Intercept	-1.1068	0.1507	3.54e-13
mortgage	-0.7319	0.1853	8.24e-05
pledge	-1.0246	0.6312	0.1053
long term	1.1420	0.2644	1.63e-05
collective	1.0166	0.2213	4.76e-06
private	0.7816	0.3234	0.0158

Table 5.6: Parametric coefficients

	edf	F	p-value
s(entry year,log loan amount)	10.64	6.357	$<$ 2e-16
s(entry year): short term	1.00	2.616	0.106
s(entry year): long term	1.00	18.747	1.6e-05
R-sq.(adj) = 0.504			n = 1379

Table 5.7: Approximate significance of smooth terms

5.5 Model Validation

5.5.1 Out-of-sample Validation

Credit scoring model can be used to measure the credit risk of exiting borrowers and predict the credit risk of future applicants. For these purposes, we are primately interested in out-of-sample validation and out-of-time validation. For this reason we first decompose the sample into training and testing sample according to a uniform distribution. The training sample with 1236 observations is for model training and the other with 143 observations used to check the performance of the model. The training sample covers 1236 commercial loans over the period between 1998 and 2004.

To evaluate the performance of the proposed credit scoring model, we will employ a Receiver Operating Characteristic(ROC) curve. The ROC curve is a graphical representation of the tradeoff between Type-I (Sensitivity) error and Type-II (Specificity) error for every possible cut off. Sensitivity is the probability of identifying a good loan among the loans that have scores indicating that they are good loans. Specificity is the probability of identifying a bad loan among the loans that

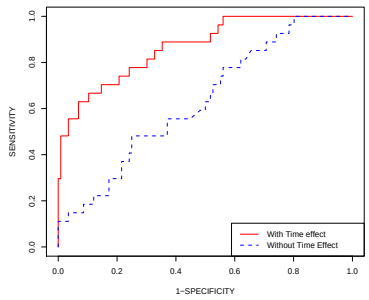


Figure 5.5: ROC curves of testing sample.

Model	*Value*
With time effect	0.8684547
Without time effect	0.6325032

Table 5.8: Area under curve

have scores indicating that they are bad loans. Sensitivity is usually placed on the Y-axis and Specificity on the X-axis. A good credit scoring model is one that has high sensitivity and specificity, which means that the ROC curve of this model climbs rapidly towards upper left corner of the graph.

To show the importance of introducing the time effect we also estimate a model with the same covariates as obtained from model selection procedure but without time varying effects. Figure 5.5 shows the comparison of ROC curves obtained by the model with time varying effect and a model without time varying effect based on the testing sample. It is easily seen that the performance of model with time effect is visually better than that of model without consideration of time effect. To be more accurate, we calculate the area under the ROC curve, which measures the speed of the rise of ROC curve to upper left corner. In case of perfect performance the area under the ROC curve is equal to 1. A comparison of the area under the ROC curves among the models is reported in Table 5.8.

5.5.2 Out-of-time Validation

Practically more important is an out-of-time validation. This means based on available data at time point t we want to score a credit starting at time point $t + 1$. To arrive this we use the current information on loan performance to estimate a model and use this to predict the performance of the loans expiring in the coming period. With time progressing, we get more information due to the expired loans and based on this new information we update our model parameters and predict again. By repeating this process, we could practically assess the credit risk in a continuous way.

We illustrate this idea by an example. Assume that we have loan performance information prior to August 31st, 2003 and want to predict the loan performance expired in September, 2003. Since we use entry year as a variable, a possible difficulty here lies in a situation that some entry years of the loans expired in September, 2003 may exceed the range of the entry years of loans prior to August 31st, 2003, So that nonparametric curves have to be extended out of range of the values their fit is based upon. We use Taylor expansion from the boundary of the smoothing function to realize this prediction. For example, for the unknown function $f(time, amount)$ we use the approximation

$$f(t_1, m_0) \approx f(t_0, m_0) + f'(t_0, m_0)(t_1 - t_0) + \frac{1}{2} f''(t_0, m_0)(t_1 - t_0)^2 \qquad (5.6)$$

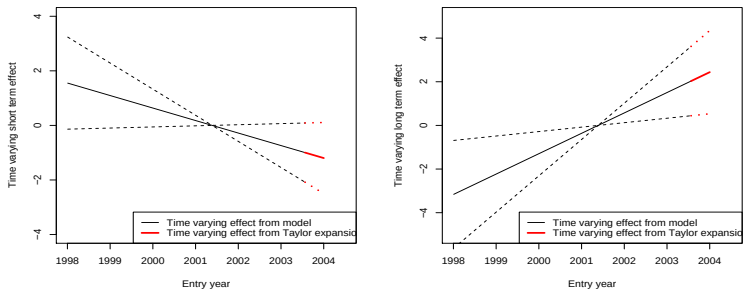
where t_1 is the period to predict, t_0 is the boundary value of function $f(time, amount)$ in $time$ direction, and m_0 is the amount, which is the same in t_1 and t_0. For the unknown time varying functions we use the same strategy shown above. We first develop a model with the same structure as before based on the data prior to August 31st, 2003. We then calculate the Taylor expansion for each part and predict the loan performance. The results are shown in Figure 5.6. To show the importance of added information, we repeat the same procedure for loans expired in July, 2004, which is shown in Figure 5.7. Comparing the short term effect and interaction in Figure 5.6 with those in Figure 5.7 we find that the effects have slightly changed, especially for interaction between the entry year and the loan amount.

5.6 Concluding Remarks

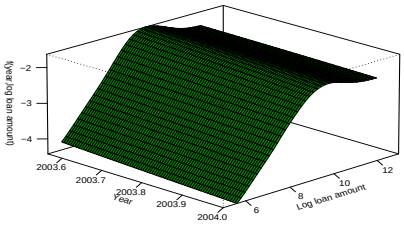
In this chapter, we show the application of a generalized logit model with time-varying effect to a credit scoring model for Chinese SMEs loan data and their estimation by P-spline. We also demonstrate that the effects of some predictors like long term maturities and loan amounts are changing over entry time. From the estimation results we can also verify the path of Chinese economy in last decade. The results of out-of-sample validation and out-of-time validation show that this consideration of varying time effect not only improves the performance of a credit scoring model, but also can detect a credit crisis earlier.

The current mortgage mess is related to the performace of the subprime mortgage or bad credit mortgage. Although a credit scoring model with time varying effect can do little to prevent the trigger of a credit crisis, it can detect a credit crisis much earlier and promote banks to react more rapidly than others if the reshuffle strategy mentioned above is applied to subprime mortgage data.

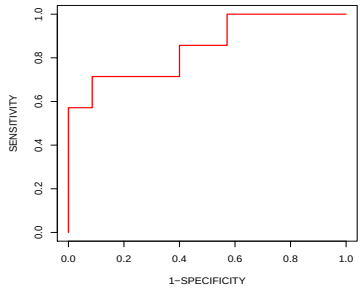
At the beginning of a credit crisis, the loans will not all suddenly become non-performing, but only few defaults happen. However, if we reshuffle a model with consideration of time varying effect or entry year effect monthly and take use of the information, the model would detect this phenomenon and change the direction of time varying effect, which is shown in Figure 5.6 and Figure 5.7 as an example. Thus, a careful check by the analyst will find that the model actually gives signals that even downgrade the loans that entry later than those default loans. A swift reaction would help a bank survive a credit crisis.



(a) Short term (b) Long term

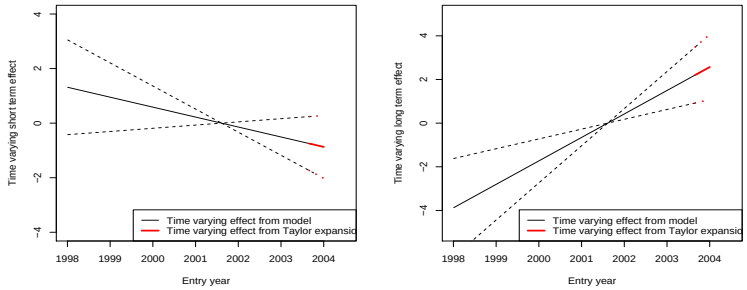


(c) Interaction between entry year and log loan amount

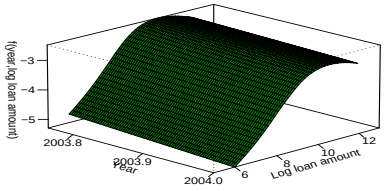


(d) ROC curve for predictions in September, 2003 (area under the curve = 0.849).

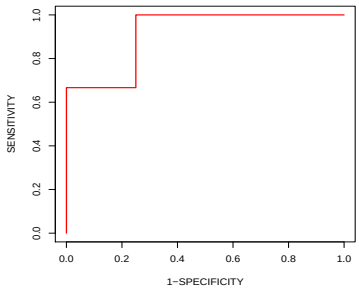
Figure 5.6: Taylor expansion and ROC curve for loans expired in September, 2003.



(a) Short term (b) Long term



(c) Interaction between entry year and log loan amount



(d) ROC curve for predictions in July, 2004 (area under the curve = 0.917).

Figure 5.7: Taylor expansion and ROC curve for loans expired in July, 2004.

Chapter 6

Summary

In this study penalized spline is proved to be a practical tool for some problems in the credit market. Due to its nonparametric nature, penalized spline is able to capture the nonlinear relationship between the economics variable as well as to detect the potential varying effect. The estimation of the unknown smooth function is relatively easy to implement because of its link to the mixed model.

In Chapter 2 we began with the introduction of penalized spline and its extension. The statistical models discussed in this chapter were linked to the issues in the credit market in later chapters. In Chapter 3, an additive model was developed to capture the nonlinear relationship between different interest rates and the prices of mortgage-backed securities. To identify the effects of different pools, grouping factors are used in estimating the unknown functions. To cope with the correlated errors, this additive model is estimated by being rewritten it as a mixed model. Based on the estimated first derivatives of the smooth functions, hedging the price changes against changes in interest rates is implemented. The out-of-sample prediction is considered by using Taylor expansion.

The impact of burnout out is investigated in Chapter 4 through an use of an additive model with a bivariate function. By analyzing the the estimated bivariate function we find that there is an interaction between burnout and scheduled factors. Moreover, the impact of burnout effect on the prices can be divided into two stages.

A generalized model consisting of a bivariate component, a varying coefficient component and a tradition linear combination component is used to explore the credit risk for small and medium-sized enterprises in China. The estimation results

identify not only the existence of the interaction between loan amount and entry year, but also the varying effects of the loan maturities. The improvement of the credit risk model is also verified by the out-of-sample and out-of-time validation. As in Chapter 3, the out-of-time prediction is implemented by the Taylor expansion. The idea of using the result of out-of-time prediction as an indicator to a credit crisis could be an interesting area for further investigation.

Appendix A: The Details of the Matrix in Model (3.1)

The details of the matrix in model (3.1) are as follows,

$$\tilde{P} = X\beta + Zu + \varepsilon$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & s_{11} & s_{11}^2 & l_{11} & l_{11}^2 \\ \vdots & \vdots & \vdots & & \\ 1 & s_{1T} & s_{1T}^2 & l_{1T} & l_{1T}^2 \\ \vdots & \vdots & \vdots & & \\ 1 & s_{n1} & s_{n1}^2 & l_{n1} & l_{n1}^2 \\ \vdots & \vdots & \vdots & & \\ 1 & s_{nT} & s_{nT}^2 & l_{nT} & l_{nT}^2 \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \end{bmatrix}, \qquad \mathbf{u} = \begin{bmatrix} U_{c_1} \\ \vdots \\ U_{c_n} \\ u_{1s} \\ \vdots \\ u_{ks} \\ u_{1l} \\ \vdots \\ u_{k_l} \end{bmatrix}$$

$$Cov \begin{bmatrix} u \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \sigma_c^2 I & 0 & 0 & 0 \\ 0 & \sigma_{u_s}^2 I & 0 & 0 \\ 0 & 0 & \sigma_{u_l}^2 I & 0 \\ 0 & 0 & 0 & \sigma_\varepsilon^2 R \end{bmatrix}$$

$$\mathbf{Z} = \begin{bmatrix} 1 & \dots & 0 & (s_{11} - \kappa_{1s})_+^2 & \dots & (s_{11} - \kappa_{k_s})_+^2 & (l_{11} - \kappa_{1l})_+^2 & \dots & (l_{11} - \kappa_{k_l})_+^2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \dots & 0 & (s_{11} - \kappa_{1s})_+^2 & \dots & (s_{11} - \kappa_{k_s})_+^2 & (l_{11} - \kappa_{1l})_+^2 & \dots & (l_{11} - \kappa_{k_l})_+^2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & (s_{n1} - \kappa_{1s})_+^2 & \dots & (s_{n1} - \kappa_{k_s})_+^2 & (l_{n1} - \kappa_{1l})_+^2 & \dots & (l_{n1} - \kappa_{k_l})_+^2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 1 & (s_{nT} - \kappa_{1s})_+^2 & \dots & (s_{nT} - \kappa_{k_s})_+^2 & (l_{nT} - \kappa_{1l})_+^2 & \dots & (l_{nT} - \kappa_{k_l})_+^2 \end{bmatrix}$$

Appendix B: Technical Details in Chapter 4

To cope with the financial data, we also assume the commonly used AR(1) structure for the correlated errors and estimate the coefficients. We replace the smooth functions $f_1(c_i)$, $f_2(s_t)$, $f_3(l_t)$ and $f_4(BO_{it}, schedule_{it})$, respectively, by high dimensional basis functions in the form

$$f_1(c_i) = B(c_i)u, \quad f_2(s_t) = \tilde{B}(s_t)\tilde{u}, \quad f_3(l_t) = \dot{B}(l_t)\dot{u}$$

where $B(c_i)$, $\tilde{B}(s_t)$ and $\dot{B}(l_t)$ are built as high dimensional basis functions built, e.g. from vectors of B-splines or truncated polynomials (see Rupper, Wand & Carroll 2003 and Wood 2006) and u, $\tilde{u}$ and $\dot{u}$ are the corresponding parameters. By analogy with the univariate case we replace $f_4(BO_{it}, schedule_{it})$ with a high dimensional basis

$$f_4(BO_{it}, schedule_{it}) = \bar{B}(BO_{it}, schedule_{it})\bar{u}$$

where $\bar{B}(BO_{it}, schedule_{it})$ is a basis in two directions. This gives a high dimensional parametric model with parameters, u, $\tilde{u}$, $\dot{u}$ and $\bar{u}$. To overcome the overfitting problem due to the high dimensionality, we extend the least square to the penalized least square by including penalty terms on the coefficients in the form, $\lambda u^T Du$, $\tilde{\lambda}\tilde{u}\tilde{D}\tilde{u}$, $\dot{\lambda}\dot{u}\dot{D}\dot{u}$ and $\bar{\lambda}\bar{u}_j^T\bar{D}_j\bar{u}_j$,

$$\begin{aligned} min \quad & \sum [p_{it} - B(c_i)u - \tilde{B}(s_t)\tilde{u} - \dot{B}(l_t)\dot{u} - \bar{B}(BO_{it}, schedule_{it})\bar{u}]^2 \\ & + \lambda u^T Du + \tilde{\lambda}\tilde{u}\tilde{D}\tilde{u} + \dot{\lambda}\dot{u}\dot{D}\dot{u} + \bar{\lambda}\bar{u}_j^T\bar{D}_j\bar{u}_j \end{aligned} \tag{6.1}$$

If we assume the coefficients u, $\tilde{u}$, $\dot{u}$ and $\bar{u}$ to be taken from normal distributions e.g. $u \sim N(0, \sigma^2 D^*)$, $\tilde{u} \sim N(0, \tilde{\sigma}^2 \tilde{D}^*)$, $\dot{u} \sim N(0, \dot{\sigma}^2 \dot{D}^*)$, $\bar{u} \sim N(0, \bar{\sigma}^2 \bar{D}^*)$, D^*, $\tilde{D}^*$, $\dot{D}^*$ and $\bar{D}^*$ as generalized inverse of D, $\tilde{D}$, $\dot{D}$ and $\bar{D}$ respectively, $\sigma^2 = 1/\lambda$, $\tilde{\sigma}^2 = 1/\tilde{\lambda}$, $\dot{\sigma}^2 = 1/\dot{\lambda}$ and $\bar{\sigma}^2 = 1/\bar{\lambda}$. The model now becomes a Generalized Linear Mixed Model (GLMM). Thus, the estimation of the smoothing parameters λ, $\tilde{\lambda}$, $\dot{\lambda}$ and $\bar{\lambda}$ and the prediction of the coefficients u, $\tilde{u}$, $\dot{u}$ and $\bar{u}$ can be obtained from standard mixed model software. The estimated price is now as follows,

$$\hat{p}_{it} = B(c_i)\hat{u} + \tilde{B}(s_t)\hat{\tilde{u}} + \dot{B}(l_t)\hat{\dot{u}} + \bar{B}(BO_{it}, schedule_{it})\hat{\bar{u}}$$

where $\hat{u}$, $\hat{\tilde{u}}$, $\hat{\dot{u}}$ and $\hat{\bar{u}}$ are solutions for minimization problem 6.1.

Appendix C: Technical Details in Chapter 5

For our investigation of (5.5) we assume that the basis components are (approximately) orthogonal, which holds exactly, if covariates x_j are set orthogonal. This simplification implies that $\boldsymbol{B}^T\boldsymbol{W}\boldsymbol{B} + \boldsymbol{D}(\lambda)$ decomposes to a block diagonal matrix with blocks of the form $B_j^T W B_j + \lambda_j D_j$, $j \in \mathcal{I}$ and where B_j is the basis $B_j(t_i), i = 1, \ldots, n$ analogously for indices $j \in \mathcal{C}\backslash\mathcal{D}$ and $j \in \mathcal{D}$. This simplification is shown useful to shed some light on our model selection routine. First, it can be shown that (see Krivobokova and Kauermann (2007))

$$\hat{\lambda}_j = \frac{\text{tr}(S_j)}{\hat{u}_j^T D_j \hat{u}_j}$$

with $S_j = B_j(B_j^T W B_j + \lambda_j D_j)^{-1} B_j W$ as j-th component smoothing matrix. Inserting $\hat{\lambda}_j$ in (5.4) allows to rewrite the maximum likelihood (5.5) to

$$l(\hat{\beta}, \hat{\lambda}) \approx l(\hat{\beta}, \hat{u}) - \frac{1}{2}\sum_j \{\text{tr}(S_j) - \log|I - S_j|\}$$

where the sum as over the index sets, $\mathcal{I}$, $\mathcal{C}\backslash\mathcal{D}$ and $\mathcal{D}$ respectively. Let now ϱ_k be the eigenvalues of $B_j^T W B_j$, $k = 1, \ldots, K$ with K as spline dimension and assume without loss of generality $D_j = I$, like for truncated polynomials. Then

$$\log|I - S_j| = \sum_{k=1}^{K} \log(1 - \frac{\varrho_k}{\varrho_k + \lambda}) = -\text{tr}(S_j) + \frac{1}{2}\text{tr}(S_j S_j) + \ldots \qquad (6.2)$$

Assuming truncated polynomials (or B-splines) it is easy to see that $\varrho_1 = \max(\varrho_k)$

= $O(n/K)$ and accordingly it follows that $\lambda = O(K)$ (justification is found in a technical report by Kauermann & Opsomer). Ignoring the remaining components in (6.2) we can now approximate

$$l_{\mathcal{M}}(\hat{\beta}, \hat{\lambda}) \approx l(\hat{\beta}, \hat{u}) - \sum_{j} \{\text{tr}(S_j(I - S_j/4))\} \tag{6.3}$$

The latter term in (6.3) serves as estimate for edf, the estimated degree of a model. Then we obtain

$$AIC = -2l(\hat{\beta}, \hat{u}) + 2edf \approx -2l_{\mathcal{M}}(\hat{\beta}, \hat{\lambda}) \tag{6.4}$$

hence, the model selection by maximizing $l_{\mathcal{M}}(\hat{\beta}, \hat{\lambda})$ is approximately equivalent to minimizing Akaike Information Criteria (AIC) (see Table 5.3, Table 5.4 and Table 5.5).

Appendix D: Sample R Codes

R is a free software, which can be used for statistical computing. For the estimation in this analysis the library **mgcv** is necessary, which can be downloaded from **http://cran.r-project.org/src/contrib /Descriptions/mgcv.html**. The variables are stored in a R data file: "data". The following R codes show the implementation of main models in all chapters and the criteria for model selection,

library(mgcv)

attach(data)

Model1<-gamm(price1 $\sim$ s(s_t)+s(l_t), method = "REML", data = impactdata, correlation = corAR1())

Model2<-gamm(price2 $\sim$ s(c_t) + s(s_t) + s(l_t)+ s($burnout_t, SF_t$), method = "REML", data = burndata, random=list(id.num= 1), correlation=corAR1())

Model3<-gamm(status $\sim$ s(year,amount)+s(year,by=short)+long+s(year,by=long) +mo+pl+col+priv, family="binomial")

n <- length(status)

mu <- Modelgamfitted.values

resi <- status - mu

A <- mu * (1-mu)

trS <- sum(Modelgamedf)

dev <- 2*sum(-log(dbinom(status, size=1, prob= mu)))

AIC <- dev + 2 * trS

Bibliography

Shanghai securities news (2006). http://www.cnstock.com/paper_new/html/2006-10/24/node_15.htm.

Aerts, M., G. Claeskens, and M. P. Wand (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference 103*, 455–470.

Aït-Sahalia, Y. and J. Duarteb (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics 116*, 9–47.

Aït-Sahalia, Y. and A. W. Lo (1998). Nonparametric estimation of state-price densities implicit in finanical asset prices. *Journal of Finance 53*, 499–547.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proc. 2nd. International Symposium on Information Theory*, 267–281.

Altman, E. I. and G. Sabato (2007). Modelling credit risk for smes: Evidence from the us market. *ABACUS 43 (3)*, 332–357.

Banz, R. and M. Miller (1978). Prices for state-contingent claims: some estimates and applications. *Journal of Business 51*, 653–672.

Beck, T. and A. Demirguc-Kunt (2006). Small and medium-size enterprises: Access to finance as a growth constraint. *Journal of Banking and Finance 30 (11)*, 2931–2943.

Behr, P. and A. Guttler (2007). Credit risk assessment and relationship lending: An empirical analysis of german small and medium-sized enterprises. *Journal of Small Business Management 45 (2)*, 194–213.

Berger, A. N. and W. S. Frame (2007). Small business credit scoring and credit availability. *Journal of Small Business Management 45 (1)*, 5–22.

Bielecki, R. T. and M. Rutkowski (2002). *Credit Risk: Modeling, Valuation and Hedging*. Springer Verlag.

Black, F. and M. Scholes (1973). The valuation of options and corporate liablilities. *Journal of Polititical Economy 8 (3)*, 637–654.

Blöchlinger, A. and M. Leippold (2006). Economic benefit of powerful credit scoring. *Journal of Banking & Finance 30*, 851–873.

Bluhm, C., L. Overbeck, and C. Wagner (2003). *An Introduction to Credit Risk Modeling*. Chapman & Hall/CRC.

Boudoukh, J., M. Richardson, R. Stanton, and R. Law (1995). A new strategy for dynamically hedging mortgage–backed securities. *Journal of Derivatives 2*, 60–77.

Boudoukh, J., M. Richardson, R. Stanton, and R. Law (1997). Pricing mortgage-backed securities in a multifactor interest rate environment: a multivariate density estimation approach. *Review of financial Studies 10*, 405–446.

Breeden, D. and R. H. Litzenberger (1978). Prices of state-contingent claims implicit in option prices. *Journal of Business 51*, 621–651.

Campbell, J. Y., A. Y. Lo, and A. C. MacKinly (1997). *Econometrics of Financial Markets*. Princeton University Press.

Charlier, E. and A. v. Bussel (2003). Prepayment behavior of dutch mortgagors: an empirical analysis. *Real Estate Economics 31*, 165–204.

Clapp, J. M., Y. S. Deng, and X. D. An (2006). Unobserved heterogeneity in models of competing mortgage termination risks. *Real Estate Economics 34(2)*, 243–273.

Clarke, G., R. Cull, M. S. M. Peria, and M. S. Sanchez (2005). Bank lending to small businesses in latin america: does bank origin matter? *Journal of Money, Credit and Banking 37 (1)*, 83–118.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik 31(4)*, 377–403.

Crook, J. N., D. B. Edelman, and L. C. Thomas (2007). Recent developments

in consumer credit risk assessment. *European Journal of Operational Research 183 (3)*, 1447–1465.

Currie, I. D. and M. Durbán (2002). Flexible smoothing with p-splines: a unified approach. *Statistical Modelling 4*, 333–349.

Deng, Y., J. M. Quigley, and R. Van Order (2000). Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica 68(2)*, 275–307.

Downing, C., R. Stanton, and N. Wallace (2005). An empirical test of a two–factor mortgage valuation model: How much do house prices matter? *Real Estate Economics 33*, 681–710.

Duffie, D. and K. J. Singleton (1999). Modeling term structures of defaultable bonds. *Review of Financial Studies 12*, 687–720.

Dunn, K. B. and J. J. McConnell (1981a). A comparison of alternativemodels for pricing gnma mortgage–backed securities. *Journal of Finance 36*, 471–483.

Dunn, K. B. and J. J. McConnell (1981b). Valuation of mortgage–backed securities. *Journal of Finance 36*, 599–617.

Dunn, K. B. and C. S. Spatt (1986). Ihe effect of refinancing costs and market imperfections on the optimal call strategy and the pricing of debt contracts. *Carnegie–Mellon University*, Working paper.

Durbán, M. and I. Currie (2003). A note on p–spline additive models with correlated errors. *Computational Statistics 18*, 263–292.

Edwards, D. and T. Havránek (1987). A fast model selection procedure for large families of models. *Journal of the American Statistical Association 82*, 205–211.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with b-splines and penalities. *Statistical Science 11*, 89–102.

Fan, J. (2005). A selective overview of nonparametric methods in financial econometrics. *Statistical Science 20(4)*, 317–337.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and its Applications*. Chapman & Hall.

Gaussel, N. and J. Tamine (2004). *Valuation of mortgage backed securities, Statistical Tools for Finance and Insurance*, Chapter 9, pp. 201–224. Springer Verlag.

Hand, D. J. (2005). Good practice in reatail credit scorecard assessment. *Journal of the Operation Research Society 56*, 1109–1117.

Hand, D. J. and W. E. Henley (1997). Statistical clasification methods in consumer credit scoring: a review. *Journal of Royal Statistical Society, Series A 160*, 523–541.

Härdle, W. and A. Yatchew (2002). Dynamic nonparametric state price density estimation using constrained least squares and the bootstrap. *Working paper SFB 373, 2002-16*, 443–459.

Harville, D. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika 61*, 383–385.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association 72*, 320–338.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall.

Hastie, T. and R. Tibshirani (1993). Varying–coefficient models (with discussion). *Journal of Royal Statistical Society, Series B 55*, 757–796.

Hayre, L. (1994). A simple statistical framework for modeling burnout and refinancing behavior. *The journal of fixed income 4(3)*, 69–74.

Hildreth, C. and J. P. Houck (1968). Some estimators for a linear model with random coefficients. *Journal of the American Statistical Association 63*, 584–595.

Hull, J. and A. White (1995). The impact of default risk on the prices of options and other derivative securities. *Journal of banking and finance 19*, 299–322.

Hurvich, C. M., J. S. Simonoff, and C. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of Royal Statistical Society, Series B. 60*, 271–293.

Hurvich, C. M. and C.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika 76(2)*, 297–307.

Hussain, J., C. Millman, and H. Matlay (2006). Sme financing in the uk and in china: a comparative perspective. *Journal of Small Business and Enterprise Development 13 (4)*, 584–599.

Jarrow, R., D. Ruppert, and Y. Yu (2004). Estimating the interest rate term structure of corporate debt with a semiparametric penalized spline model. *Journal of the American Statistical Association 99*, 57–66.

Jarrow, R. and S. Turnbull (1995). Pricing derivatives on financial securities subject to credit risk. *Journal of Finance 50(1)*, 53–85.

Jegadrrsh, N. and X. Ju (2000). A non–parametric prepayment model and valuation of mortgage–backed securities. *The Journal of Fixed Income June*, 50–67.

Johnston, E. and L. V. Drunen (1988). Pricing mortgage pools with heterogeneous mortgagors: Empirical evidence. *working paper University of Utah*.

Kariya, T., F. Ushiyama, and S. R. Pliska (2002). A 3-factor valuation model for mortgage-backed securities (mbs). *Working paper No.543*. Institute of Economic Research, Kyoto University.

Kau, J. B. and D. C. Keenan (1995). An overview of the option–theoretic pricing of mortgages. *Journal of Housing Research 6(2)*, 217–244.

Kau, J. B., D. C. Keenan, W. J. Muller, and J. F. Epperson (1992). A generalized valuation model for fixed-rate residential mortgages. *Journal of Money, Credit and Banking 24*, 279–299.

Kau, J. B., D. C. Keenan, W. J. Muller, and J. F. Epperson (1995). The valuation at origination of fixed rate mortgages with default and prepayment. *Journal of Real Estate Finance and Economics 11*, 5–39.

Kau, J. B. and T. M. Springer (1992). The prepayment option on mortgage securities: A random coefficient approach. *Review of Quantitative Finance and Accounting 2*, 33–45.

Kau, J. B. and T. M. Springer (1993). An analysis of financial and nonfinancial prepayment of gnma securities with a varying coefficient model. *Journal of*

Real Estate Research 8, 69–86.

Kauermann, G. (2006). Nonparametric models and their estimation. *Allgemeines Statistisches Archiv 90*, 135–150.

Kawasaki, Y. and T. Ando (2005). Estimation term structure using nonlinear splines: a penalized likelihood approach. *MODSIM 2005 International Congress on Modelling and Simulation*, 864–870.

Krivobokova, T. and G. Kauermann (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Assosiation 102*, 1328–1337.

Krivobokova, T., G. Kauermann, and T. Archontakis (2006). Estimating the term structure of interest rates using penalized splines. *Statistical Papers 47(3)*, 443–459.

LaCour-Little, M. and R. K. Green (2002). Parameter stability and the valuation of mortgages and mortgage–backed securities. *University of Wisconsin's Department of Real Estate and Urban Land Economics Working Paper*.

LaCoure-Little, M., M. Marshoun, and C. Maxam (1999). Estimation prepayments on fixed–rate mortgages: A multi-variate kernel regression approach using loan–level data. *Freddie Mac Working Paper*.

Lagnado, R. and S. Osher (1997). A technique for calibrating derivative security pricing models: numerical solution of the inverse problem. *Journal of Computational Finance 1*, 13–25.

Leonard, K. J. (1992). Credit scoring models for the evaluation of small business loan applications. *IMA Journal of Mathematics Applied in Business & Industry 4*, 89–95.

Levin, A. (2001). Active passive decomposition in burnout modeling. *The journal of fixed income 10*, 27–40.

Li, B. and D. Mehta (2001). Restructuring of chinese banking industry. *China & World Economy November 3*.

Liu, W. and J. Cela (2006). Improving credit scoring by generalized additive model. SAS Global Forum 2007.

Mallows, C. L. (1973). Some comments on cp. *Technometrics 15(4)*, 661–675.

Mattey, J. and N. Wallace (1998). Housing prices and the (in)stability of mortgage prepayment models: Evidence from california. *Federal Reserve Bank of San Francisco Working Paper 98–05*.

Mattey, J. and N. Wallace (2001). Housing–price cycles and prepayment rate of u.s. mortgage pools. *Journal of Real Estate Finance and Economics 23*, 161–184.

Maxam, C. and M. LaCour-Little (2001). Applied nonparametric regression techniques: Estimation prepayments on fixed–rate mortgage–backed securities. *Journal of Real Estate Finance and Economics 23*, 139–160.

Merton, R. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science 4*, 141–183.

Merton, R. (1974). On the pricing of corporate debt : the risk structure of interest rates. *Journal of Finance 2 (2)*, 449–470.

Müller, M. and B. Rönz (2000). *Credit scoring using semiparametric methods, Measuring Risk in Complex Stochastic Systems*, Chapter 5, pp. 85–102. Springer Verlag.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications 9(1)*, 141–142.

Ngo, L. and M. P. Wand (2004). Smoothing with mixed model software. *Journal of Statistical Software 9(1)*.

Opsomer, J., Y. Wang, and Y. Yang (2001). Nonparametric regression with correlated errors. *Statistical Science 16*, 134–153.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science 1*, 502–518.

Phillips, R. A. and J. H. Vanderhoff (2004). The conditional probability of foreclosure: An empirical analysis of conventional mortgage loan defaults. *Real Estate Economics 32 (4)*, 571–587.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-Plus*. Springer Verlag.

Popova, L., E. Popova, and E. George (2006). Bayesian forecasting of prepayment rates for individual pools of mortgages. *Working Paper*.

Rau, R., J. Gampe, H. P. Eilers, and B. D. Marx (2007). Modeling seasonal effects on the lexis surface. *2007 PAA Annual Meeting*.

Rebonato, R. (1998). *Interest Rate Option Models*. John Wiley & Sons.

Richard, S. and R. Roll (1989). Prepayment on fixed-rate mortgage-backed securities. *Journal of Portfolio Management 15*, 73–82.

Robinson, G. K. (1991). That blup is a good thing: the estimation of random effects. *Statistical Science 6*, 15–51.

Rubinstein, M. and J. Jackwerth (1996). Recovering probability distributions from option prices. *Journal of Finance 51*, 1611–1631.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics 11*, 735–757.

Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge University Press.

Schönbucher, P. (2003). *Credit Derivatives Pricing Models*. John Wiley & Sons.

Schwartz, E. S. and W. N. Torous (1989). Prepayment and the valuation of mortgage–backed securities. *Journal of Finance 44*, 375–392.

Schwartz, E. S. and W. N. Torous (1992). Prepayment, default, and the valuation of mortgage pass-through securities. *The Journal of Business 65(2)*, 221–239.

Schwartz, E. S. and W. N. Torous (1993). Mortgage prepayment and default decision: A poisson regression approach. *Journal of the American Real Estate and Urban Economics Association 21(4)*, 431–449.

Spahr, E. R. and M. A. Sundermann (1992). The effect of prepayment mortgage–backed securities. *Journal of Housing Research 3(2)*, 381–400.

Stanton, R. (1995). Rational prepayment and the valuation of mortgage-backed securities. *Review of Financial Studies 8*, 677–708.

Stanton, R. (1996). Unobservable heterogeneity and rational learning: pool–specific versus generic mortgage bakced security prices. *Journal of Real Estate Finance and Economics 12(3)*, 243–263.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B 36(2)*, 111–147.

Swamy, P. (1970). Effcient inference in a random coecient regressionmodel. *Econometrics 38*, 311–323.

Timmis, G. C. (1985). Valuation of gnma mortgage–backed securities with transaction costs, heterogeneous households and endogenously generated prepayment rates. *Carnegie–Mellon University Working paper*.

Tomas, L. C., R. W. Oliver, and D. J. Hand (2005). A survey of the issues in consumer credit modelling research. *Journal of The Operational Research Society 56*, 1006–1015.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A, 26.*, 359–372.

Wegener, M. and G. Kauermann (2008). Examining heterogeneity in implied equity risk premium using penalized splines. *Advances in Statistical Analysis 92(1)*, 35–56.

Wood, S. N. (2000). Modelling and smoothing parameter selection with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B 62(2)*, 413–428.

Wood, S. N. (2006). *Generalized Additive Models*. Chapman & Hall/CRC.